

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-261072

(43)Date of publication of application : 03.10.1997

(51)Int.Cl.

H03M 7/40

(21)Application number : 08-063573

(71)Applicant : FUJITSU LTD

(22)Date of filing : 19.03.1996

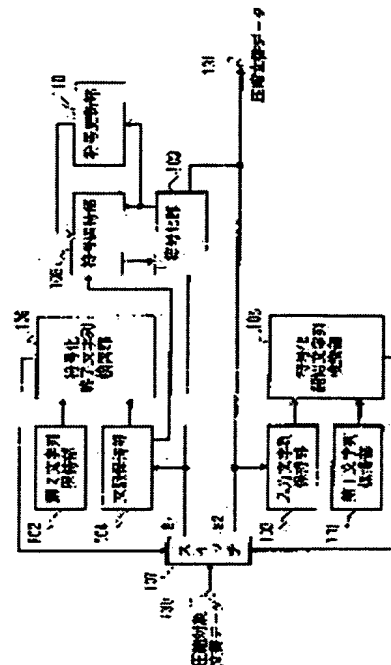
(72)Inventor : MURASHITA KIMITAKA
YOSHIDA SHIGERU
OKADA YOSHIYUKI

(54) DOCUMENT MANAGING DEVICE, DATA COMPRESSING METHOD AND DATA RESTORING METHOD

(57)Abstract:

PROBLEM TO BE SOLVED: To obtain a document managing device for preparing compressed document data capable of retrieving a key word by providing a controlled character string storing means, a coding means coding an inputted character and a retrieving means, etc., retrieving a starting controlled character string and a finishing controlled character string from an inputted character string.

SOLUTION: The device is provided with a controlled character string storing means 109 storing respectively at least one starting controlled character string and finishing controlled character string, a coding part 109 coding the inputted character and retrieving parts 105 and 106 retrieving the starting controlled character string and the finishing controlled character string from an inputted character string. Then when the retrieving means 105 and 106 retrieve the starting controlled character string, the processing of outputting coded data obtained by coding inputted character strings after this by means of the coding part 109 as the element of compressed document data is started. When the retrieving parts 105 and 106 retrieve the finish controlled character string, the inputted character strings are outputted as the element of compressed document data as it is without coding by means of the coding part 109.



LEGAL STATUS

[Date of request for examination] 24.03.2000

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 3305191

[Date of registration] 10.05.2002

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-261072

(43) 公開日 平成9年(1997)10月3日

(51) Int.Cl.⁶

H 0 3 M 7/40

識別記号

庁内整理番号

9382-5K

F I

H 0 3 M 7/40

技術表示箇所

審査請求 未請求 請求項の数18 O L (全 35 頁)

(21) 出願番号 特願平8-63573

(22) 出願日 平成8年(1996)3月19日

(71) 出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中4丁目1番
1号

(72) 発明者 村下 君孝

神奈川県川崎市中原区上小田中1015番地富
士通株式会社内

(72) 発明者 吉田 茂

神奈川県川崎市中原区上小田中1015番地富
士通株式会社内

(72) 発明者 岡田 佳之

神奈川県川崎市中原区上小田中1015番地富
士通株式会社内

(74) 代理人 弁理士 遠山 勉 (外1名)

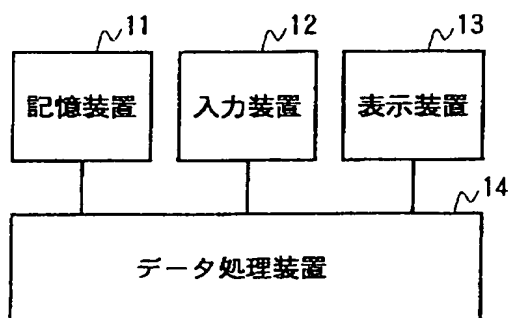
(54) 【発明の名称】 文書管理装置及びデータ圧縮方法及びデータ復元方法

(57) 【要約】

【課題】 キーワード検索が行える圧縮文書ファイルを作成する文書管理装置を提供する。

【解決手段】 圧縮すべき文書データが与えられたときに、その文書データ中に終了制御文字列が現れるまで、各文字をそのまま圧縮文書ファイル内に書き込む処理と、文書データ中に開始制御文字列が現れるまで、各文字を符号化したデータを圧縮文書ファイルに書き込む処理とが交互に実行されるように文書管理装置を構成する。また、圧縮文書ファイルを復元する際には、圧縮文書ファイル内のデータ中に終了制御文字列が現れるまで、各データ(文字)をそのまま出力する処理と、圧縮文書ファイル内のデータの復元結果に開始制御文字列が現れるまで、復号を行う処理とが交互に実行されるようにする。

第1実施形態の文書管理装置のブロック図



【特許請求の範囲】

【請求項1】 入力された文字列に応じた圧縮文書データを作成する文書管理装置において、

1個以上の開始制御文字列と1個以上の終了制御文字列を記憶する制御文字列記憶手段と、

入力された文字を符号化した符号化データを出力する符号化手段と、

入力文字列から開始制御文字列及び終了制御文字列を検索する検索手段と、

前記検索手段によって前記開始制御文字列が検索されたときに、以降の入力文字列を前記符号化手段によって符号化した符号化データを圧縮文書データの要素として出力する処理を開始し、前記検索手段によって前記終了制御文字列が検索されたときには、前記符号化手段による符号化を行わずに、以降の入力文字列をそのまま圧縮文書データの要素として出力する処理を開始する制御手段とを備えることを特徴とする文書管理装置。

【請求項2】 入力された圧縮文書データを復元した文書データを出力する文書管理装置において、1個以上の開始制御文字列と1個以上の終了制御文字列を記憶する制御文字列記憶手段と、

入力された符号を複合した文字を出力する復号手段と、復元を終えた文書データの末尾に開始制御文字列あるいは終了制御文字列が存在するかどうかを判別する判別手段と、

この判別手段によって開始制御文字列の存在が判別されたときに、以降の圧縮文書データを前記復号手段によって復号した文字を文書データの要素として出力する処理を開始し、前記判別手段によって終了制御文字列が検索されたときには、前記復号手段による復号を行わずに、以降の圧縮文書データをそのまま文書データの要素として出力する処理を開始する制御手段とを備えることを特徴とする文書管理装置。

【請求項3】 前記符号化手段は、動的モデルを用いて、前記文字に対応する符号を出力し、前記制御手段は、前記検索手段によって前記終了制御文字列が検索されたときに、前記符号化手段が用いる動的モデルを初期化することを特徴とする請求項2記載の文書管理装置。

【請求項4】 前記制御手段は、以降の入力文字列を非符号化データとして出力する処理を開始する際に、前記検索手段によって検索された終了制御文字列を圧縮文書データの要素として出力することを特徴とする請求項2または請求項3に記載の文書管理装置。

【請求項5】 前記制御手段は、前記検索手段によって前記終了制御文字列が検索されたときには、前記符号化手段による符号化を行わずに、以降の入力文字列を、入力文字と出力文字との対応関係が定められた置換表を用いて置換し、置換結果を非符号化データとして出力する処理を開始することを特徴とする請求項2ないし請求項

4のいずれかに記載の文書管理装置。

【請求項6】 さらに、圧縮文書データに対してある文字列の検索が指示された際に、その文字列を前記置換表を用いて置換する置換手段と、

この置換手段によって置換された文字列を用いた検索を実行する検索手段とを備えることを特徴とする請求項5記載の文書管理装置。

【請求項7】 幾つかの文書要素の前後に、それぞれ、その文書要素の内容に応じた開始制御文字列と終了制御文字列が挿入された文書データを対象とする文書管理装置であって、

データを表示するための表示手段と、

1個以上の開始制御文字列と1個以上の終了制御文字列を記憶する制御文字列記憶手段と、

圧縮すべき文書データ内の文字を順に読み出す第1読出手段と、

この第1読出手段によって読み出された文字をそのまま圧縮文書ファイルの要素として出力するとともに、その文字をインデックスファイルの要素として出力する第1出力手段と、

前記第1読出手段によって前記制御文字列記憶手段内のいずれかの開始制御文字列と同じ文字列が読み出されたときに前記第1読出手段の動作を中止させる第1制御手段と、

この第1制御手段によって前記第1読出手段の動作が中止されたときに、前記文書データ内の文字の読み出しを開始する第2読出手段と、

この第2読出手段によって読み出された文字に対応する符号を、圧縮文書データの要素として出力する第2出力手段と、

前記第2読出手段によって前記制御文字列記憶手段内のいずれかの終了制御文字列と同じ文字列が読み出されたときに、前記第2読出手段の動作を中止させるとともに、前記第1読出手段の動作を再開させる第2制御手段と、

前記圧縮文書ファイルと前記インデックスファイルを記憶する記憶手段と、

所定の指示が与えられた際に、前記記憶手段に記憶されたインデックスファイル内の、前記終了制御文字列で区切られた各データをインデックスとして前記表示手段に表示する表示制御手段と、

この表示制御手段によって表示されたインデックスの中から1つのインデックスを指定する指定手段と、

この指定手段によって指定されたインデックスの前記圧縮文書ファイル内での格納位置を特定する格納位置特定手段と、

前記圧縮文書ファイル内の、前記格納位置特定手段で特定された格納位置以降のデータを前記制御文字列記憶手段に記憶されているいずれかの終了制御文字列が復元されるまで復元する部分復元手段とを備えることを特徴と

する文書管理装置。

【請求項8】 さらに、前記第1出力手段が出力を開始する度に、圧縮文書ファイルの要素としてそれまでに出力されたデータの積算サイズを検出して記憶する積算サイズ検出記憶手段を備え、

前記格納位置特定手段は、前記積算サイズ検出記憶手段によって記憶されている積算サイズに基づき、前記インデックスの圧縮文書ファイル内での格納位置を特定することを特徴とする請求項7記載の文書管理装置。

【請求項9】 前記部分復元手段は、

前記圧縮文書ファイル内の、前記格納位置特定手段で特定された格納位置以前のデータを処理済のデータであると認識する復元不要データ認識手段と、

圧縮文書ファイル内の未処理のデータを1文字分ずつ順に読み出す第1データ読出手段と、

この第1データ読出手段によって読み出されたデータを復号結果として出力する第1復号手段と、

この第1復号手段によって前記制御文字列記憶手段内のいずれかの開始制御文字列と同じ文字列が出力されたときに、前記第1データ読出手段の動作を中止させる第1読出制御手段と、

この第1読出制御手段によって前記第1データ読出手段の動作が中止されたときに、前記圧縮文書ファイル内の未処理のデータの読み出しを開始する第2データ読出手段と、

この第2データ読出手段によって読み出されたデータを復号した文字を出力する第2復号手段と、

この第2復号手段によって前記制御文字列記憶手段内のいずれかの終了制御文字列と同じ文字列が出力されたときに、前記第2データ読出手段の動作を中止させる第2読出制御手段と、

この第2読出制御手段による制御が行われたときに、前記第2データ読出手段が読み出した文字列が前記特定手段によって特定されたインデックスの末尾に含まれる開始制御文字列に対応する終了制御文字列でなかった場合には、前記第1データ読出手段の動作を再開させる第3読出制御手段とを備えることを特徴とする請求項7または請求項8記載の文書管理装置。

【請求項10】 幾つかの文書要素の前後に、それぞれ、その文書要素の内容に応じた開始制御文字列と終了制御文字列が挿入された文書データを対象とする文書管理装置であって、

データを表示するための表示手段と、

1個以上の開始制御文字列と1個以上の終了制御文字列を記憶する制御文字列記憶手段と、

圧縮すべき文書データ内の文字を順に読み出す第1読出手段と、

この第1読出手段によって読み出された文字を静的符号化した符号を、圧縮文書ファイルの要素として出力するとともに、その文字をインデックスファイルの要素とし、

て出力する第1出力手段と、

前記第1読出手段によって前記制御文字列記憶手段内のいずれかの開始制御文字列と同じ文字列が読み出されたときに前記第1読出手段の動作を中止させる第1制御手段と、

この第1制御手段によって前記第1読出手段の動作が中止されたときに、前記文書データ内の文字の読み出しを開始する第2読出手段と、

この第2読出手段によって読み出された文字を動的符号化した符号を、圧縮文書ファイルの要素として出力する第2出力手段と、

前記第2読出手段によって前記制御文字列記憶手段内のいずれかの終了制御文字列と同じ文字列が読み出されたときに、前記第2読出手段の動作を中止させ、前記第2出力手段が動的符号化に用いるモデルを初期化し、前記第1読出手段の動作を再開させる第2制御手段と、

前記第1出力手段が出力を開始する度に、前記第1出力手段及び第2出力手段がそれまでに圧縮文書ファイルの要素として出力したデータの積算サイズを検出し、記憶する積算サイズ検出記憶手段と、

前記圧縮文書ファイルと前記インデックスファイルとを記憶する記憶手段と、

所定の指示が与えられた際に、前記記憶手段に記憶されているインデックスファイル内の、前記開始制御文字列で区切られたデータをそれぞれインデックスとして前記表示手段に表示する第1表示制御手段と、

この表示制御手段によって表示されたインデックスの中から1つのインデックスを指定する指定手段と、

前記積算サイズ検出記憶手段内に記憶されている積算サイズに基づき、前記指定手段によって指定されたインデックスの前記圧縮文書ファイル内での格納位置を特定し、前記圧縮文書ファイル内のそのインデックス以前のデータを処理済のデータであると認識する復元不要データ認識手段と、

圧縮文書ファイル内の未処理のデータを読み出す第1データ読出手段と、

この第1データ読出手段によって読み出されたデータを静的復号した文字を出力する第1復号手段と、

この第1復号手段によって前記制御文字列記憶手段内のいずれかの開始制御文字列と同じ文字列が復号されたときに、前記第1データ読出手段の動作を中止させる第1復号制御手段と、

この第1復号制御手段によって前記第1データ読出手段の動作が中止されたときに、前記圧縮文書ファイル内の未処理のデータの読み出しを開始する第2データ読出手段と、

この第2データ読出手段によって読み出されたデータを動的復号した文字を出力する第2復号手段と、

この第2復号手段によって前記制御文字列記憶手段内のいずれかの終了制御文字列と同じ文字列が復号されたと

きに、前記第2データ読出手段の動作を中止させるとともに前記第2復号手段が動的復号に用いるモデルを初期化する第2復号制御手段と、

この第2復号制御手段による制御が行われたときに、前記第2復号手段によって復号された文字列が、前記指定手段によって指定されたインデックスの末尾に含まれる開始制御文字列に対応する終了制御文字列でなかった場合に、前記第1読出手段の動作を再開させる第3復号制御手段と、を備えることを特徴とする文書管理装置。

【請求項11】 開始制御文字列と終了制御文字列が挿入された原データを圧縮するデータ圧縮方法であって、前記原データから開始制御文字列及び終了制御文字列を検索する検索ステップと、前記検索ステップにおいて前記開始制御文字列が検索されたときに、以降の原データを符号化した符号化データを圧縮データの要素として出力する処理を開始し、前記検索ステップにおいて前記終了制御文字列が検索されたときには、符号化を行わずに、以降の原データをそのまま圧縮データの要素として出力する処理を開始するデータ処理ステップとを備えることを特徴とするデータ圧縮方法。

【請求項12】 前記データ処理ステップは、動的モデルを用いて符号化を行い、前記検索ステップにおいて前記終了制御文字列が検索されたときには、その動的モデルを初期化することを特徴とする請求項11記載のデータ圧縮方法。

【請求項13】 前記データ処理ステップは、以降の原データを符号化した符号化データを圧縮データの要素として出力する処理を開始する際に、前記検索ステップにおいて検索された前記終了制御文字列を圧縮データの要素として出力する請求項11または請求項12記載のデータ圧縮方法。

【請求項14】 前記データ処理ステップは、前記検索ステップにおいて前記終了制御文字列が検索されたときには、以降の原データを所定の置換表を用いて置換したデータを圧縮データの要素として出力する処理を開始することを特徴とする請求項11または請求項12記載のデータ圧縮方法。

【請求項15】 開始制御文字列をその末尾に有するデータと、終了制御文字列をその末尾に有するデータを符号化したデータとが混在する圧縮データを復元するデータ復元方法であって、復元を終えたデータの末尾に開始制御文字列あるいは終了制御文字列が存在するか否かを判別する判別ステップと、この判別ステップにおいて開始制御文字列の存在が判別されたときに、以降の圧縮データを復号した文字を復元結果として出力する処理を開始し、前記判別ステップにおいて終了制御文字列が検索されたときには、以降の圧縮データをそのまま復元結果として出力する処理を開始

するデータ処理ステップとを備えるデータ復元方法。

【請求項16】 前記データ処理ステップは、動的モデルを用いて復号を行い、前記検索ステップにおいて前記終了制御文字列が検索されたときには、その動的モデルを初期化することを特徴とする請求項15記載のデータ復元方法。

【請求項17】 前記データ処理ステップは、復号した文字を出力する処理を開始する際に、最初に復号される終了制御文字列を復元結果として取り扱わないことを特徴とする請求項15または請求項16記載のデータ復元方法。

【請求項18】 前記データ処理ステップは、前記検索ステップにおいて前記終了制御文字列が検索されたときには、以降の圧縮データを所定の置換表を用いて置換したデータを復元結果として出力する処理を開始することを特徴とする請求項15または請求項16記載のデータ復元方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文書管理装置およびデータ圧縮方法およびデータ復元方法に関し、特に、文書データを圧縮して管理する文書管理装置と、文書データなどを圧縮・復元する際に用いるデータ圧縮方法及びデータ復元方法に関する。

【0002】

【従来の技術】近年、文字コード、ベクトル情報、画像情報など様々な種類のデータがコンピュータで扱われるようになっている。また、扱われるデータ量も急激に増大してきており、伝送時間を短縮するためや、記憶装置を効率的に利用するために、データを圧縮することが行われている。

【0003】たとえば、アーカイバと呼ばれるアプリケーションでは、1つ以上のファイルから、1つの圧縮データファイルが作成される。アーカイバを用いて、使用頻度の低いファイルや古いファイルなどを圧縮することによって、ファイル容量を削減することができる。そして、ファイルの内容を通信によって授受する際に、アーカイバによって作成された圧縮データファイルを用いれば、通信に要する時間が短縮され、通信コストも低減する。

【0004】また、ハードディスクやフロッピーディスクなどのドライブを圧縮ドライブとして動作させることも行われている。圧縮ドライブを有するシステムでは、ユーザがファイルの書き込みを指示した場合、そのファイルが自動的に圧縮されて圧縮ドライブ内に格納される。そして、ユーザがファイルの読み出しを指示した場合には、圧縮ドライブ内のファイルが自動的に復元される。

【0005】なお、コンピュータシステムで扱われるデータには、文字、機械語、画像、音声など様々なものがあるので、上述のようなファイル圧縮の際には、各種の

データに適用可能な符号化方式であるユニバーサル符号化方式が用いられている。具体的には、データ（文字）の再現性を利用した辞書型符号化方式や、確率統計型符号化方式に分類される算術符号化方式、Splay-Tree符号化方式などが用いられている。

【0006】

【発明が解決しようとする課題】さて、圧縮されていないファイルに対しては、キーワード検索を行うことにより、その内容を確認することができる。たとえば、SGML (Standard Generalized Markup Language)形式による文書データでは、文書データ中の特定の要素の前後に、その要素の内容に応じたタグが使用されている。このため、SGML形式の文書データでは、そのファイルの中から目的とする情報に付けられているタグを検索し、その後に記憶されている文字列を読み出してやれば、必要な情報を得ることができる。

【0007】しかしながら、SGML形式の文書データを圧縮した場合、タグの検索が行えなくなってしまう。このため、タイトルだけを確認したい場合にも、圧縮ファイル全体を復元しなければならず、確認作業に時間がかかっていた。

【0008】そこで、本発明の課題は、キーワード検索が行える圧縮文書データを作成する文書管理装置を提供することにある。また、本発明の他の課題は、キーワード検索が行える圧縮データを作成するデータ圧縮方法と、そのデータ圧縮方法によって作成された圧縮データを復元するデータ復元方法を提供することにある。

【0009】

【課題を解決するための手段】本発明の文書管理装置は、幾つかの文書要素の前後に、それぞれ、その文書要素の内容に応じた開始制御文字列と終了制御文字列が挿入された文書データを対象とする。

【0010】本発明の第1の文書管理装置は、1個以上の開始制御文字列と1個以上の終了制御文字列を記憶する制御文字列記憶手段と、入力された文字を符号化した符号化データを出力する符号化手段と、入力文字列から開始制御文字列及び終了制御文字列を検索する検索手段と、検索手段によって開始制御文字列が検索されたときに、以降の入力文字列を符号化手段によって符号化した符号化データを圧縮文書データの要素として出力する処理を開始し、検索手段によって終了制御文字列が検索されたときには、符号化手段による符号化を行わずに、以降の入力文字列をそのまま圧縮文書データの要素として出力する処理を開始する制御手段とを備える。

【0011】すなわち、第1の文書管理装置は、文書データに基づき、圧縮されていないデータと圧縮データとが混在する圧縮文書データを作成する。従って、第1の文書管理装置が作成する圧縮文書データは、復元しなくとも、キーワード検索を行うことによってその内容を確認できる。このため、第1の文書管理装置によれば、効

率的な文書データ管理が行えることになる。

【0012】なお、第1の文書管理装置によって作成された圧縮文書データは、1個以上の開始制御文字列と1個以上の終了制御文字列を記憶する制御文字列記憶手段と、入力された符号を複合した文字を出力する復号手段と、復元を終えた文書データの末尾に開始制御文字列あるいは終了制御文字列が存在するか否かを判別する判別手段と、この判別手段によって開始制御文字列の存在が判別されたときに、以降の圧縮文書データを復号手段によって復号した文字を文書データの要素として出力する処理を開始し、判別手段によって終了制御文字列が検索されたときには、復号手段による復号を行わずに、以降の圧縮文書データをそのまま文書データの要素として出力する処理を開始する制御手段とを備える文書管理装置によって復元される。

【0013】本発明の第1の文書管理装置では、符号化手段として、動的モデル（例えば、ダイナミックハフマン）を用いて文字に対応する符号を出力する手段を採用するとともに、制御手段として、検索手段によって終了制御文字列が検索されたときに、符号化手段が用いる動的モデルを初期化する手段を採用することができる。このように文書管理装置を構成した場合には、その内容の一部だけを復元することができる圧縮文書データが作成されることになる。

【0014】また、第1の文書管理装置では、制御手段として、以降の入力文字列を非符号化データとして出力する処理を開始する際に、検索手段によって検索された終了制御文字列を圧縮文書データの要素として出力する手段を採用することもできる。

【0015】このように文書管理装置を構成した場合には、文書データ内に存在していた開始制御文字列と終了制御文字列で挟まれた文書要素が、そのまま記憶された圧縮文書データが作成される。このため、この文書管理装置によれば、圧縮文書データに対する、キーワード検索がさらに容易に行えることになる。

【0016】本発明の第1の文書管理装置では、制御手段として、検索手段によって終了制御文字列が検索されたときには、符号化手段による符号化を行わずに、以降の入力文字列を、入力文字と出力文字との対応関係が定められた置換表を用いて置換し、置換結果を非符号化データとして出力する処理を開始する手段を採用することができる。

【0017】このように文書管理装置を構成した場合、そのまま読みとれるデータが含まれていない圧縮文書データが作成される。従って、この文書管理装置が作成した圧縮文書データを、インターネットを用いて転送したとしても、中間のマシンによってその内容が読みとれることがない。このため、この文書管理装置によれば、データ通信時の秘匿性を高めることができる。

【0018】なお、文字を置換して出力するよう装置を

構成する場合には、文書管理装置に、圧縮文書データに対してある文字列の検索が指示された際に、その文字列を置換表を用いて置換する置換手段と、この置換手段によって置換された文字列を用いた検索を実行する検索手段とを付加することが望ましい。

【0019】本発明の第2の文書管理装置は、データを表示するための表示手段と、1個以上の開始制御文字列と1個以上の終了制御文字列を記憶する制御文字列記憶手段と、圧縮すべき文書データ内の文字を順に読み出す第1読出手段と、この第1読出手段によって読み出された文字を圧縮文書ファイルの要素として出力するとともに、その文字をインデックスファイルの要素として出力する第1出力手段と、第1読出手段によって制御文字列記憶手段内のいずれかの開始制御文字列と同じ文字列が読み出されたときに第1読出手段の動作を中止させる第1制御手段と、この第1制御手段によって第1読出手段の動作が中止されたときに、文書データ内の文字の読み出しを開始する第2読出手段と、この第2読出手段によって読み出された文字に対応する符号を、圧縮文書データの要素として出力する第2出力手段と、第2読出手段によって制御文字列記憶手段内のいずれかの終了制御文字列と同じ文字列が読み出されたときに、第2読出手段の動作を中止させるとともに、第1読出手段の動作を再開させる第2制御手段と、圧縮文書ファイルとインデックスファイルを記憶する記憶手段と、所定の指示が与えられた際に、記憶手段に記憶されたインデックスファイル内の、開始制御文字列で区切られた各データをインデックスとして表示手段に表示する表示制御手段と、この表示制御手段によって表示されたインデックスの中から1つのインデックスを指定する指定手段と、この指定手段によって指定されたインデックスの圧縮文書ファイル内の格納位置を特定する格納位置特定手段と、圧縮文書ファイル内の、格納位置特定手段で特定された格納位置以降のデータを制御文字列記憶手段に記憶されているいずれかの終了制御文字列が復元されるまで復元する部分復元手段とを備える。

【0020】すなわち、本発明の第2の文書管理装置は、文書データに基づき、圧縮されていないデータ（第1出力手段が出力するデータ）と圧縮データ（第2出力手段が出力するデータ）とが混在する圧縮文書ファイルを作成するとともに、第1出力手段が出力するデータからなるインデックスファイルを作成する。

【0021】記憶手段に記憶されたインデックスファイルの内容は、表示制御手段によって、たとえば、CRTなどの表示手段に表示される。ユーザは、キーボードやマウスといった入力装置から構成される指定手段を用いて、表示手段に表示された複数のインデックスの中から、1つのインデックスを指定する。また、格納位置特定手段は、たとえば、指定されたインデックスを圧縮文書ファイル内で検索することによって、そのインデック

スの格納位置を特定する。そして、部分復元手段は、圧縮文書ファイル内の、その格納位置以降のデータを制御文字列記憶手段に記憶されているいずれかの終了制御文字列が復元されるまで復号する。

【0022】このように、第2の文書管理装置では、圧縮文書ファイルの内容を一部分だけ復元する機能が設けられているので、圧縮文書ファイル全体を復元しなくとも、その内容を確認できる。このため、第2の文書管理装置によれば、ハードディスク装置などによって構成される記憶手段の記憶容量を有効に利用しつつ、効率的な文書データ処理が行えることになる。

【0023】この第2の文書管理装置に、第1出力手段が出力を開始する度に、圧縮文書ファイルの要素としてそれまでに出力されたデータの積算サイズを検出して記憶する積算サイズ検出記憶手段を付加し、格納位置特定手段として、積算サイズ検出記憶手段によって記憶されている積算サイズに基づき、インデックスの圧縮文書ファイル内の格納位置を特定する手段を用いることもできる。

【0024】また、第2の文書管理装置では、部分復元手段として、圧縮文書ファイル内の、格納位置特定手段で特定された格納位置以前のデータを処理済のデータであると認識する復元不要データ認識手段と、圧縮文書ファイル内の未処理のデータを1文字分ずつ順に読み出す第1データ読出手段と、この第1データ読出手段によって読み出されたデータを復号結果として出力する第1復号手段と、この第1復号手段によって制御文字列記憶手段内のいずれかの開始制御文字列と同じ文字列が出力されたときに、第1データ読出手段の動作を中止させる第1読出制御手段と、この第1読出制御手段によって第1データ読出手段の動作が中止されたときに、圧縮文書ファイル内の未処理のデータの読み出しを開始する第2データ読出手段と、この第2データ読出手段によって読み出されたデータを復号した文字を出力する第2復号手段と、この第2復号手段によって制御文字列記憶手段内のいずれかの終了制御文字列と同じ文字列が出力されたときに、第2データ読出手段の動作を中止させる第2読出制御手段と、この第2読出制御手段による制御が行われたときに、第2データ読出手段が読み出した文字列が特定手段によって特定されたインデックスの末尾に含まれる開始制御文字列に対応する終了制御文字列でなかった場合には、第1データ読出手段の動作を再開させる第3読出制御手段とからなる手段を用いることができる。

【0025】このような構成の部分復元手段を用いた場合には、指定手段で指定したインデックスに応じた範囲のデータを復元させることができることになる。本発明の第3の文書管理装置は、データを表示するための表示手段と、1個以上の開始制御文字列と1個以上の終了制御文字列を記憶する制御文字列記憶手段と、圧縮すべき文書データ内の文字を順に読み出す第1読出手段と、こ

の第1読出手段によって読み出された文字を静的符号化した符号を、圧縮文書ファイルの要素として出力するとともに、その文字をインデックスファイルの要素として出力する第1出力手段と、第1読出手段によって制御文字列記憶手段内のいずれかの開始制御文字列と同じ文字列が読み出されたときに第1読出手段の動作を中止させる第1制御手段と、この第1制御手段によって第1読出手段の動作が中止されたときに、文書データ内の文字の読み出しを開始する第2読出手段と、この第2読出手段によって読み出された文字を動的符号化した符号を、圧縮文書ファイルの要素として出力する第2出力手段と、第2読出手段によって制御文字列記憶手段内のいずれかの終了制御文字列と同じ文字列が読み出されたときに、第2読出手段の動作を中止させ、第2出力手段が動的符号化に用いるモデルを初期化し、第1読出手段の動作を再開させる第2制御手段と、第1出力手段が出力を開始する度に、第1出力手段及び第2出力手段がそれまでに圧縮文書ファイルの要素として出力したデータの積算サイズを検出し、記憶する積算サイズ検出記憶手段と、圧縮文書ファイルとインデックスファイルとを記憶する記憶手段と、所定の指示が与えられた際に、記憶手段に記憶されているインデックスファイル内の、開始制御文字列で区切られたデータをそれぞれインデックスとして表示手段に表示する第1表示制御手段と、この表示制御手段によって表示されたインデックスの中から1つのインデックスを指定する指定手段と、積算サイズ検出記憶手段内に記憶されている積算サイズに基づき、指定手段によって指定されたインデックスの圧縮文書ファイル内での格納位置を特定し、圧縮文書ファイル内のそのインデックス以前のデータを処理済のデータであると認識する復号不要データ認識手段と、圧縮文書ファイル内の未処理のデータを読み出す第1データ読出手段と、この第1データ読出手段によって読み出されたデータを静的復号した文字を出力する第1復号手段と、この第1復号手段によって制御文字列記憶手段内のいずれかの開始制御文字列と同じ文字列が復号されたときに、第1データ読出手段の動作を中止させる第1復号制御手段と、この第1復号制御手段によって第1データ読出手段の動作が中止されたときに、圧縮文書ファイル内の未処理のデータの読み出しを開始する第2データ読出手段と、この第2データ読出手段によって読み出されたデータを動的復号した文字を出力する第2復号手段と、この第2復号手段によって制御文字列記憶手段内のいずれかの終了制御文字列と同じ文字列が復号されたときに、第2データ読出手段の動作を中止させるとともに第2復号手段が動的復号に用いるモデルを初期化する第2復号制御手段と、この第2復号制御手段による制御が行われたときに、第2復号手段によって復号された文字列が、指定手段によって指定されたインデックスの末尾に含まれる開始制御文字列に対応する終了制御文字列でなかった場合に、第1読

出手段の動作を再開させる第3復号制御手段とを備える。

【0026】すなわち、本発明の第3の文書管理装置では、文書データに基づき、静的符号化により圧縮されたデータ（第1出力手段が出力するデータ）と動的符号化により圧縮されたデータ（第2出力手段が出力するデータ）とが混在する圧縮文書ファイルが作成されるとともに、第1出力手段が出力する圧縮文書データに対応する非圧縮データからなるインデックスファイルが作成される。

【0027】記憶手段に記憶されたインデックスファイルの内容は、表示制御手段によって、たとえば、CRTなどの表示手段に表示される。ユーザは、キーボードやマウスといった入力装置から構成される指定手段を用いて、表示手段に表示された複数のインデックスの中から、1つのインデックスを指定する。

【0028】復号不要データ認識手段は、検出記憶手段内の積算サイズに基づき、ユーザによって指定されたインデックスの圧縮文書ファイル内での格納位置を特定し、そのインデックス以前のデータを処理済のデータであると認識する。そして、この復号不要データ認識手段によって処理済であると認識されたデータ以降のデータに対して、ユーザによって指定されたインデックスの末尾に含まれる開始制御文字列に対応する終了制御文字列が復元されるまで、各部による処理が繰り返される。

【0029】このように、第3の文書管理装置では、2種類の圧縮方法を用いて文書データを圧縮した圧縮文書ファイルが作成されるので、圧縮文書ファイルのサイズが小さく、ハードディスク装置などによって構成される記憶手段の記憶容量を有効に利用できることになる。また、キーワード検索可能なインデックスファイルが作成されるので、圧縮文書ファイルを復元しなくとも、その内容を推定できる。また、圧縮文書ファイルの内容を一部分だけ復元する機能が設けられているので、必要な部分だけを復元することができる。このため、第3の文書管理装置によれば、効率的な文書データ処理が行えることになる。

【0030】本発明のデータ圧縮方法は、幾つかのデータ要素の前後に、それぞれ、終了制御文字列と開始制御文字列が挿入された原データを対象とする。本発明のデータ圧縮方法は、原データから開始制御文字列及び終了制御文字列を検索する検索ステップと、検索ステップにおいて開始制御文字列が検索されたときに、以降の原データを符号化した符号化データを圧縮データの要素として出力する処理を開始し、検索ステップにおいて終了制御文字列が検索されたときには、符号化を行わずに、以降の原データをそのまま圧縮データの要素として出力する処理を開始するデータ処理ステップとを備える。

【0031】このように、本発明のデータ圧縮方法では、圧縮されていないデータと圧縮データとが混在する

圧縮データ、すなわち、キーワード検索可能な圧縮データが作成される。

【0032】このデータ圧縮方法によって作成された圧縮データファイルは、以下に記すデータ復元方法によって復元される。本発明のデータ復元方法は、復元を終えたデータの末尾に開始制御文字列あるいは終了制御文字列が存在するか否かを判別する判別ステップと、この判別ステップにおいて開始制御文字列の存在が判別されたときに、以降の圧縮データを復号した文字を復元結果として出力する処理を開始し、判別ステップにおいて終了制御文字列が検索されたときには、以降の圧縮データをそのまま復元結果として出力する処理を開始するデータ処理ステップとを備える。

【0033】本発明のデータ圧縮方法では、データ処理ステップとして、動的モデルを用いて符号化を行い、検索ステップにおいて終了制御文字列が検索されたときには、その動的モデルを初期化するステップを用いることができる。

【0034】このデータ圧縮方法によって作成された圧縮データを復元する際には、上述のデータ復元方法のデータ処理ステップとして、動的モデルを用いて復号を行い、検索ステップにおいて終了制御文字列が検索されたときには、その動的モデルを初期化するステップを用いる。

【0035】本発明のデータ圧縮方法では、データ処理ステップとして、以降の原データを符号化した符号化データを圧縮データの要素として出力する処理を開始する際に、検索ステップにおいて検索された終了制御文字列を圧縮データの要素として出力するステップを採用することもできる。

【0036】このデータ圧縮方法によって作成された圧縮データを復元する際には、本発明のデータ復元方法におけるデータ処理ステップとして、復号した文字を出力する処理を開始する際に、最初に復号される終了制御文字列を復元結果として取り扱わないステップを採用する。

【0037】また、本発明のデータ圧縮方法では、データ処理ステップとして、検索ステップにおいて終了制御文字列が検索されたときには、以降の原データを所定の置換表を用いて置換したデータを圧縮データの要素として出力する処理を開始するステップを用いることもできる。

【0038】このデータ圧縮方法によって作成された圧縮データを復元する際には、上述のデータ復元方法のデータ処理ステップとして、検索ステップにおいて終了制御文字列が検索されたときには、以降の圧縮データを所定の置換表を用いて置換したデータを復元結果として出力する処理を開始するステップを採用する。

【0039】

【発明の実施の形態】以下、本発明を図面を用いて詳細

に説明する。まず、本発明の文書管理装置が対象とする文書データの記述形式の概要を説明する。本発明の文書管理装置は、文書を制御する文字と文書とが同一のデータ内に格納されている文書データを対象とする。ここでは、SGML形式の文書データを対象とした場合を例に、実施形態の文書管理装置を説明する。SGML形式で記述された文書データのようなは、1986年にISOが制定した文書形式の国際規格である。SGML形式による文書データでは、文書データ中の特定の要素の前後に、その要素の内容に応じたタグと呼ばれる制御文字列が使用される。たとえば、文書のタイトルを表す要素の前には、“<TITLE>”といった開始タグが使用され、その要素の後には、“</TITLE>”といった終了タグが使用される。

【0040】第1実施形態

第1実施形態の文書管理装置は、文書データをファイル化する際に、圧縮データと非圧縮データが混在するファイル（以下、圧縮文書ファイルと表記する）を作成する。

【0041】図1に、本発明の第1実施形態による文書管理装置の構成を示す。図示したように、第1実施形態の文書管理装置は、記憶装置11と入力装置12と表示装置13とデータ処理装置14とを備える。記憶装置11は、いわゆる、磁気ディスク記憶装置であり、圧縮文書ファイルなどを記憶する。入力装置12は、キーボード及びマウスとその周辺機器から構成されている。表示装置13は、CRT(Cathod Ray Tube)とその周辺機器からなり、記憶装置11内に記憶された圧縮文書ファイルの復元結果などを表示するために用いられる。

【0042】データ処理装置14は、CPU(Central Processing Unit)を中心として構成されており、文書データの編集機能を有する。また、データ処理装置14は、入力装置11から与えられる指示に応じて、文書データから圧縮文書ファイルを作成する処理や、圧縮文書ファイルを文書データに復元する処理を実行する。

【0043】以下、本文書管理装置（データ処理装置14）の動作を説明する。まず、図2に示した機能ブロック図を用いて、データ処理装置14による圧縮文書ファイルの作成動作を説明する。

【0044】図示したように、データ処理装置14は、スイッチ107と、スイッチ107のS2端子側に設けられた入力文字列保持部103と第1文字列保持部101と符号化開始文字列検索部105と、スイッチ107のS1端子側に設けられた第2文字列保持部102と文脈保持部104と符号化終了文字列検索部106と符号保持部108と符号化部109と符号更新部110とからなる。

【0045】圧縮すべき文書データは、文字毎に、入力端子130からスイッチ107に供給される。スイッチ107は、入力された文字を、S1端子あるいはS2端

子のいずれか一方の端子から出力するスイッチである。スイッチ107は、圧縮文書ファイルの作成開始時、S2端子から文字を出力する。

【0046】まず、スイッチ107が、文字をS2端子側に出力しているときの各部の動作を説明する。スイッチ107のS2端子から文字が出力されている場合、入力文字列保持部103と第1文字列保持部101と符号化開始文字列検索部105が機能する。S2端子からの文字は、出力端子131から出力されて、圧縮文書ファイルの構成データとされるとともに、入力文字列保持部103に10 入力される。入力文字列保持部103は、所定値N1個の文字からなる文字列を保持する能力を有し、S2端子から供給される文字で、保持する文字列の内容を更新する。すなわち、入力文字列保持部103は、M(<N1)個の文字からなる文字列を保持していた場合に、S2端子から文字が供給された際には、その文字列の末尾に供給された文字を追加する。また、N1個の文字からなる文字列を保持していた場合に、S2端子から文字が供給された際には、その文字列の先頭から1文字を取り除き、その末尾にS2端子からの文字を追加す 20 る。

【0047】第1文字列保持部101は、終了タグから選択された幾つかの符号化開始文字列(</SECTION>、</SUBSECTION>等)を保持している。なお、入力文字列保持部103が保持する文字列の文字数の最大値N1は、この第1文字列保持部103内の最長の符号化開始文字列の文字数となっている。

【0048】符号化開始文字列検索部105は、入力文字列保持部103に新たな文字が入力される度に、入力文字列保持部103内の文字列の末尾に、第1文字列保持部101内に保持されているいずれかの符号化開始文字列と一致する文字列が存在しているか否かを検索する。そして、いずれかの符号化開始文字列と一致する文字列が存在していなかった場合、符号化開始文字列検索部105は、何も行わず、次の文字の入力を待機する。一方、符号化開始文字列と一致する文字列が存在していた場合、符号化開始文字列検索部105は、スイッチ107のデータの出力先をS2端子からS1端子に切り替える。 30

【0049】たとえば、入力文字列保持部103内に“****</SECTION>”とい文字列が保持されているときに、S2端子から文字“>”が供給された場合、入その文字列は“****</SECTION>”に更新される。従って、符号化開始文字列検索部105は、入力文字列保持部103内の文字列の末尾に符号化開始文字列“</SECTION>”を見いだし、スイッチ107対してデータの出力先の切換を指示することになる。出力端子131からは、この時点まで、非圧縮データが出力されることになる。

【0050】次に、スイッチ107のS1端子から文字 50

が出力された場合の動作を説明する。この場合、第2文字列保持部102と文脈保持部104と符号化終了文字列検索部106と符号保持部108と符号化部109と符号更新部110が機能する。

【0051】第2文字列保持部102と文脈保持部104と符号化終了文字列検索部106は、それぞれ、第1文字列保持部101と入力文字列保持部103と符号化開始文字列検索部105と類似の動作をする。

【0052】すなわち、第2文字列保持部102は、終了タグから選択された幾つかの符号化終了文字列(<SECTION>、<SUBSECTION>等)を保持する。文脈保持部104は、第2文字列保持部102が保持する最長の符号化終了文字列と同じ長さの文字列を保持する能力を有し、S1端子から供給される文字で、内部に保持している文字列の内容を更新する。また、文脈保持部104は、保持している文字列のうち、末尾側の所定数の文字からなる文字列(文脈)を、符号保持部108に供給する。

【0053】符号化終了文字列検索部106は、文脈保持部104に新たな文字が入力される度に、文脈保持部104内の文字列の末尾に、第2文字列保持部102内に保持されているいずれかの符号化終了文字列と一致する文字列が存在しているか否かを判断する。そして、一致する文字列が存在していなかった場合、符号化終了文字列検索部106は、何も行わず、次の文字の入力を待機する。一方、いずれかの符号化終了文字列と一致する文字列が存在していた場合、符号化終了文字列検索部106は、スイッチ107のデータの出力先をS2端子からS1端子に切り替える。

【0054】符号保持部108、符号化部109、符号更新部110は、S1端子からの順次供給される文字を動的に符号化する。各部は、次のように動作する。符号保持部108は、符号化に使用する符号表を文脈毎に保持しており、文脈保持部104から通知される文脈に応じた符号表を参照・更新対象とする。符号化部109は、符号保持部108によって参照・更新対象とされた符号表を用いて、S1端子から入力された文字に対応する符号を決定し、決定した符号(圧縮データ)を出力端子131から出力する。この圧縮データの出力は、S2端子側にスイッチ107が切り替えられるまでの間、続けられる。符号更新部110は、文字の符号化が終わった際に、その文字の出現頻度が増加したことが文字と符号との対応関係に反映されるように、符号化に使用された符号表の内容を更新する。

【0055】以下、図3ないし図5を用いて、第1実施形態の文書管理装置の圧縮文書ファイルの作成手順をさらに詳細に説明する。これらの図のうち、図3は、データ処理装置14による圧縮文書ファイルの作成手順を示した流れ図である。また、図4は、本装置の圧縮対象となる文書データの一例を示した図である。図5は、図4

に示した文書データに基づき、本文書管理装置によって作成される圧縮文書ファイルの概要を示した図である。なお、以下の説明では、符号化開始文字列として、“</SECTION>”と“</SUBSECTION>”が、符号化終了文字列として、“<SECTION>”と“<SUBSECTION>”が設定されているものとする。

【0056】圧縮文書ファイルの作成は、文書データを構成する各文字をそのまま出力する非圧縮データ出力処理ループと、各文字を圧縮して出力する圧縮データ出力処理ループを交互に繰り返すことによって進められる。図3に示したように、文書データの圧縮を指示された際、データ処理装置14内では、非圧縮データ出力処理ループ（ステップS101～S103）が実行される。

【0057】非圧縮データ出力処理ループでは、まず、文書データ内の1文字（対象文字）がそのまま（図2のS2側から）出力され、圧縮文書ファイルに書き込まれる（ステップS101）。次いで、文書データを構成する全ての文字に対する処理が終了しているか否かが判断される（ステップS102）。そして、全ての文字に対する処理が終了していなかった場合（ステップS102；N）には、そのときまでに処理された文字列が、いずれかの符号化開始文字列と一致しているか否かが判断される（ステップS103）。

【0058】処理された文字列が各符号化開始文字列と一致していなかった場合（ステップS103；N）には、ステップS101からの処理が再度実行される。一方、そのときまでに処理された文字列が、符号化開始文字列の1つと一致していた場合（ステップS103；Y）には、圧縮データ出力処理ループ（ステップS104～S107）が開始される。

【0059】たとえば、図4に示した文書データに関する圧縮文書ファイルの作成を行った場合、最初に現れる符号化開始文字列は、“</SECTION>”（2行目）である。このため、文書データの先頭から2行目の“</SECTION>”までの各文字は、そのまま出力されて、圧縮文書ファイル内に記憶される。この結果、圧縮文書ファイルの先頭には、図5に示したように、文書データと同じ内容のデータが記憶されることになる。そして、“</SECTION>”の次の文字から圧縮データ出力処理が開始される。

【0060】図3に戻って、圧縮文書ファイル作成処理の説明を続ける。圧縮データ出力処理ループでは、S1側で文書データから次の1文字が読み込まれ、その対象文字に対応する符号が符号化部から出力される（ステップS104）。このステップにおける符号出力は、対象文字の文脈を参照した形で進められる。その後、符号化に用いた文脈に関する符号表の内容が更新される（ステップS105）。

【0061】次に、文書データを構成する全ての文字に

対する処理が終了しているか否かが判断され、終了していない場合（ステップS106；N）には、符号化を終えた幾つかの文字からなる文字列が、いずれかの符号化終了文字列と一致するか否かが判断される（ステップS107）。

【0062】符号化を終えた幾つかの文字からなる文字列が、全ての符号化終了文字列と一致していなかった場合（ステップS107；N）には、ステップS104からの処理が再度実行される。一方、符号化を終えた幾つかの文字からなる文字列が、符号化終了文字列の1つと一致していた場合（ステップS107；Y）には、非圧縮データ出力処理ループ（ステップS101～S103）が再度開始される。

【0063】たとえば、図4に示した文書データでは、3行目以降に最初に現れる符号化終了文字列は、“<SECTION>”（4行目）である。このため、3行目の始めから、4行目の“<SECTION>”までの各文字は符号化されて出力される。その結果、この部分の各文字は、図5の3行目に示したように、圧縮データとして圧縮文書ファイル内に格納されることになる。そして、“<SECTION>”の次の文字からの文章（2．特許請求の範囲</SECTION>…）に対して、再度、非圧縮データ出力処理ループ、圧縮データ出力処理ループによる処理が繰り返され、結局、符号化終了文字列と符号化開始文字列として指定しておいた制御文字列で挟まれた部分だけが非圧縮であり、その他の部分（<PARAGRAPH>、<TT>といった他の制御文字列を含む）が圧縮された圧縮文書ファイルが作成されていく。

【0064】この圧縮文書ファイル作成処理は、非圧縮データ出力処理ループにおいて全データに関する処理が終了した際（ステップS102；Y）、あるいは、圧縮データ出力処理ループにおいて全データに関する処理が終了した際（ステップS106；Y）に、完了する。

【0065】以下、図6に示した機能ブロック図を用いて、第1実施形態の文書管理装置（データ処理装置14）による圧縮文書ファイルの復元動作を説明する。圧縮文書ファイルを構成するデータは、入力端子230からスイッチ207に供給されている。スイッチ207は、入力された文字を、S1端子あるいはS2端子のいずれか一方の端子から出力する。

【0066】以下、スイッチ207が、データをS2端子側に供給しているときの各部の動作を説明する。なお、圧縮文書ファイルの復元は、スイッチ207のS2端子からデータが出力される状態で開始される。

【0067】スイッチ207のS2端子からデータが供給されている場合、入力文字列保持部203と第1文字列保持部201と復号開始文字列検索部205が機能する。スイッチ207のS2端子からのデータは、出力端子231から、文書データ中の1文字として出力される

とともに、入力文字列保持部203に供給されている。

【0068】入力文字列保持部203は、最大N1文字分の文字列を保持し、S2端子から供給される文字で、内部に保持する文字列の内容を更新する。第1文字列保持部201は、第1文字列保持部101が保持する符号化開始文字列と同じ文字列(</SECTION>、</SUBSECTION>等)を復号開始文字列として保持している。復号開始文字列検索部205は、入力文字列保持部203に新たなデータ(文字)が入力される度に、入力文字列保持部203内の文字列の末尾に、第1文字列保持部201内に保持されているいずれかの復号開始文字列と一致する文字列が存在しているか否かを判断する。そして、復号開始文字列と一致する文字列が存在していなかった場合、復号開始文字列検索部205は、何も行わず、次のデータの inputs を待機する。一方、復号開始文字列の1つと一致する文字列が存在していた場合、以降の文字列が圧縮されたものなので、複合処理が必要となる。このため、復号開始文字列検索部205は、スイッチ207のデータ出力先をS2からS1に切り替える。

【0069】次に、スイッチ207のS1端子からデータ(符号)が出力されるとき動作を説明する。この場合、符号保持部208と復号部209と符号更新部210と、第2文字列保持部202と文脈保持部204と符号化終了文字列検索部206が機能を開始する。

【0070】符号保持部208、復号部209、符号更新部210は、S1端子からの順次供給されるデータ(符号)を適応的に復号する。各部は、次のように動作する。符号保持部208は、復号に使用する符号表を文脈毎に保持しており、後述する文脈保持部204から通知される文脈に応じた符号表を参照・更新対象とする。復号部209は、符号保持部208によって参照・更新対象とされた符号表を用いてS1端子から入力される符号を復号する。そして、復号結果である文字を、出力端子231と文脈保持部204に供給する。符号更新部210は、復号部209による復号が行われた後に、復号結果である文字の出現頻度が増加したことが文字と符号との対応関係に反映されるように、復号に使用された符号表の内容を更新する。

【0071】文脈保持部204は、N2文字分の文字列を保持する能力を有し、復号部209から供給される文字で、保持する文字列の内容を更新する。また、文脈保持部204は、保持する文字列のうち、末尾側の所定数の文字からなる文字列を文脈として符号保持部208に供給する。第2文字列保持部202は、第2文字列保持部102(図2)が保持する符号化終了文字列と同じ文字列を、復号終了文字列として保持している。なお、N2は、第2文字列保持部202内の最長の復号終了文字列の文字数となっている。

【0072】復号終了文字列検索部206は、文脈保持

部204に新たな文字が入力される度に、文脈保持部204内の文字列の末尾に、第2文字列保持部202内に保持されているいずれかの復号終了文字列と一致する文字列が存在しているか否かを判断する。そして、復号終了文字列と一致する文字列が存在していなかった場合、復号終了文字列検索部206は、何も行わず、次の復号結果の inputs を待機する。一方、復号終了文字列と一致する文字列が存在していた場合、その後続く文字列は、圧縮されていない文字列であるので、復号終了文字列検索部206は、スイッチ207のデータの出力先をS1端子からS2端子に切り替える。

【0073】以下、図7と、圧縮文書ファイルの作成手順の説明に用いた図4および図5を参照して、第1実施形態の文書管理装置の圧縮文書ファイルの復元手順をさらに詳細に説明する。なお、図7は、データ処理装置14による圧縮文書ファイルの復元手順を示した流れ図である。

【0074】図7に示したように、圧縮文書ファイルの復元を最初に指示された際、データ処理装置14内では、非圧縮データ処理ループ(ステップS201~S203)が実行される。非圧縮データ処理ループでは、まず、圧縮文書ファイル内の最初の1文字分のデータがそのまま復元結果として出力される(ステップS201)。次いで、圧縮文書ファイル内の全てのデータに対する処理が終了しているか否かが判断される(ステップS202)。そして、全てのデータに対する処理が終了していなかった場合(ステップS202;N)には、出力を終えた幾つかの文字からなる文字列が、いずれかの復号開始文字列と一致しているか否かが判断される(ステップS203)。

【0075】出力を終えた幾つかの文字からなる文字列が各復号開始文字列と一致していなかった場合(ステップS203;N)には、ステップS201からの処理が再度実行される。一方、出力を終えた幾つかの文字からなる文字列が、復号開始文字列の1つと一致していた場合(ステップS203;Y)には、圧縮データ処理ループ(ステップS204~S207)が開始される。

【0076】たとえば、図5に示した圧縮文書ファイルが処理対象であった場合、非圧縮データ処理ループにおいて最初に見い出される復号開始文字列は、“</SECTION>”(2行目)である。このため、“</SECTION>”までの各文字は、そのまま出力され、図4の先頭2行のデータが生成される。そして、“</SECTION>”の次のデータから圧縮データ処理ループによる処理が開始されることになる。

【0077】図7に戻って、圧縮文書ファイルの復元処理の説明を続ける。圧縮データ処理ループでは、圧縮文書ファイルのデータ(符号)が必要量読み込まれ、その符号の復号結果である文字が出力される(ステップS204)。なお、復号は、既に復号を終えた文字列(文

脈)を参照した形で行われる。そして、その後、復号に用いた文脈に関する符号表の内容が更新される(ステップS205)。

【0078】次に、圧縮文書ファイル内の全てのデータに対する処理が終了しているか否かが判断される(ステップS206)。そして、全てのデータに対する処理が終了していない場合(ステップS206;N)には、復号を終えた幾つかの文字からなる文字列が、いずれかの復号終了文字列と一致するか否かが判断される(ステップS207)。

【0079】復号を終えた幾つかの文字からなる文字列が各復号終了文字列と一致していなかった場合(ステップS207;N)には、ステップS204からの処理が再度実行される。一方、復号を終えた幾つかの文字からなる文字列が復号終了文字列の1つと一致していた場合(ステップS207;Y)には、非圧縮データ処理ループ(ステップS201~S203)が再度実行される。

【0080】たとえば、図5の3行目からの圧縮データを順次復号していくと、いずれ、“<SECTION>”という文字列が復元されることになる。データ処理装置14は、このように復号終了文字列の1つと一致する文字列が復元されたときに、圧縮データ処理ループを抜けだし、非圧縮データ処理ループを開始する。

【0081】なお、データ処理装置14は、非圧縮データ処理ループにおいて全データに関する処理が終了した際(ステップS202;Y)、あるいは、圧縮データ処理ループにおいて全データに関する処理が終了した際(ステップS206;Y)に、圧縮文書ファイル復元処理を終える。

【0082】以上詳細に説明したように、第1実施形態の文書管理装置では、文書データに基づき、その内容の一部がそのままの形で記憶された圧縮文書ファイルが作成される。すなわち、キーワード検索が可能な圧縮文書ファイルが作成される。このため、本文書管理装置では、圧縮文書ファイルを復元することなく、圧縮文書ファイルの内容を推定(確認)することができる。

【0083】なお、第1実施形態の文書管理装置は、SGML形式の文書データを対象とする装置として構成してあるが、本装置は、内部に記憶させておく制御文字列を変更するだけで、他形式のデータ(文書データに限らない)を対象とする装置になる。また、当然、制御文字列ではなく制御文字を使用することも可能である。

【0084】さて、第1実施形態の文書管理装置が管理する圧縮文書ファイルに対して、タグ単位での検索ではなく、タグの構成要素である“<”や“>”の検索を実行した場合、圧縮データ内の符号が検索されてしまう場合も考えられる。このような誤った検索が行われるのを防ぐために、文書管理装置に、検索した文字の次に、非文字コードが存在していた場合には、その文字を無視し、さらに検索を続行する検索機能を持たせても良い。

また、この検索機構をさらに確実に動作させるために、圧縮文書ファイルを構成する圧縮データ内に“0x3c”(“<”のASCIIコード)、“0x3e”(“>”のASCIIコード)が現れる場合、その後に例えば“0x00”といったASCIIコードではない特定のコードが挿入されるようにしておくこともできる。なお、このように文書管理装置を構成する場合には、圧縮文書ファイルの復元時に、その特定のコードが取り除かれるようにする。

10 【0085】第2実施形態

第1実施形態の文書管理装置は、非圧縮データとして、文書データ内のデータをそのまま使った圧縮文書ファイルを作成する装置であった。これに対して、第2実施形態の文書管理装置は、文書データ内のデータそのままではなく、そのデータを所定の規則に従って置換したデータを格納した圧縮文書ファイルを作成する。すなわち、第2の書管理装置は、そのまま読みとれるデータが含まれていない圧縮文書データを作成する。第2実施形態の文書管理装置の動作手順は、第1実施形態の文書管理装置の動作手順と類似しているので、ここでは、動作内容が異なる部分だけを説明することにする。

【0086】まず、図8および図9を用いて、第2実施形態の文書管理装置による、圧縮文書ファイル作成手順を説明する。なお、図8は、第2実施形態の文書管理装置による圧縮文書ファイル作成手順を説明するための機能ブロック図であり、図9は、圧縮文書ファイル作成手順を示した流れ図である。

【0087】図8に示したように、第2実施形態の文書管理装置では、スイッチ107のS2端子からのデータ(非圧縮対象文字)は、置換部122に供給され、置換部122の出力が圧縮文書ファイル内に格納される。

【0088】置換部122には、文字と置換後の文字を対応づけた置換表を保持する置換表保持部123が接続されている。置換部122は、その置換表においてS2端子からの文字に対応づけられている文字を出力する。

【0089】すなわち、第2実施形態の文書管理装置では、図9に示したように、非圧縮データ出力処理ループ(ステップS301~S303)において、文字を出力する際には、文書データ内の文字を置換して出力(ステップS301)する。

【0090】この第2実施形態の文書管理装置によって作成された圧縮文書ファイルには、そのまま読みとれるデータは存在しない。例えば、インターネットでは、複数のマシン間でリレー式にファイル転送が行われるが、この圧縮文書ファイル形態で文書データを転送すれば、中間のマシンによってファイルの内容が読みとられることを防ぐことができる。

【0091】なお、第2実施形態の文書管理装置は、圧縮文書ファイルのキーワード検索を指示した際、そのキーワードを置換表を用いて置換したキーワードによる検

素が実行されるように構成されている。

【0092】次に、図10および図11を用いて、第2実施形態の文書管理装置による、圧縮文書ファイルの復元手順を説明する。図10は、第2実施形態の文書管理装置の圧縮文書ファイル復元手順を説明するための機能ブロック図であり、図11は、第2実施形態の文書管理装置の圧縮文書ファイル復元手順を示した流れ図である。

【0093】図10に示したように、第2実施形態の文書管理装置では、スイッチ107のS2端子からのデータ（文字）は、逆置換部222に供給され、逆置換部222の出力が圧縮文書ファイルを復元した文書データに加えられる。

【0094】逆置換部222には、置換表保持部123内の置換表に対応する逆置換表を保持する逆置換表保持部223が接続されている。逆置換部222は、その逆置換表によって、S2端子からの文字に対応づけられている文字を出力する。

【0095】すなわち、第2実施形態の文書管理装置では、図11に示したように、非圧縮データ出力処理ループ（ステップS401～S403）において、圧縮文書ファイル内のデータ（文字）を逆置換した文字を出力（ステップS401）する。

【0096】第3実施形態

第3実施形態の文書管理装置は、第1実施形態の文書管理装置を基にして構成されている。ただし、第3実施形態の文書管理装置では、非圧縮データと圧縮データが混在する圧縮文書ファイルが作成される際には、非圧縮データだけからなるインデックスファイルも作成される。また、圧縮文書ファイルの形態も第1実施形態の文書管理装置で作成される圧縮文書ファイルとは異なったものとなっている。さらに、第3実施形態の文書管理装置では、インデックスファイルを利用して復元を行う部分を指定できるようになっている。

【0097】まず、図12を用いて、第3実施形態の文書管理装置（データ処理装置）による圧縮文書ファイル作成手順を説明する。文書データの圧縮を最初に指示された際、データ処理装置内では、非圧縮データ出力処理ループ（ステップS501～S503）が開始される。非圧縮データ出力処理ループでは、まず、文書データ内の1文字（対象文字）がそのまま出力され、圧縮文書ファイルとインデックスファイルに書き込まれる（ステップS501）。次いで、文書データを構成する全ての文字に対する処理が終了しているか否かが判断される（ステップS502）。そして、処理すべき文字が残っていた場合（ステップS502；N）には、処理を終えた幾つかの文字からなり、そのときに処理された文字を含む文字列が、予め定められている符号化開始文字列の1つと一致しているかが判断される（ステップS503）。

【0098】処理を終えた文字列と一致する符号化開始文字列がなかった場合（ステップS503；N）には、ステップS501からの処理が再度実行される。一方、符号化開始文字列の1つと一致する文字列が処理されていた場合（ステップS503；Y）には、圧縮データ出力処理ループ（ステップS504～S507）が開始される。

【0099】圧縮データ出力処理ループでは、文書データから次の1文字が読み込まれ、その対象文字に対応する符号が出力される（ステップS504）。このステップにおける符号出力は、対象文字の文脈を参照した形で進められる。その後、符号化に用いた文脈に関する符号表の内容が更新される（ステップS505）。

【0100】次に、文書データを構成する全ての文字に対する処理が終了しているか否かが判断される（ステップS506）。処理すべき文字が残っていた場合（ステップS506；N）には、そのときに処理した文字を含む処理済の文字列が、予め定められている符号化終了文字列の1つと一致するか否かが判断される（ステップS507）。そして、処理した文字列が各符号化終了文字列と一致していなかった場合（ステップS507；N）には、ステップS504からの処理が再度実行される。

【0101】一方、処理した文字列が符号化終了文字列の1つと一致していた場合（ステップS507；Y）、符号表の初期化が行われる（ステップS508）。その後、ステップS507で検出した符号化終了文字列が、圧縮文書ファイルとインデックスファイルに出力され（ステップS509）、非圧縮データ出力処理ループ（ステップS501～S503）が再度開始される。

【0102】この圧縮文書ファイル作成処理は、非圧縮データ出力処理ループにおいて全データに関する処理が終了したことが検出された際（ステップS502；Y）、あるいは、圧縮データ出力処理ループにおいて全データに関する処理が終了したことが検出された際（ステップS506；Y）に、終了される。

【0103】以下、図4に示した文書データを対象とした場合を例に、第3実施形態の文書管理装置による圧縮文書ファイル作成手順をさらに具体的に説明する。なお、以下の説明では、符号化開始文字列として、“</SECTION>”と“</SUBSECTION>”が、符号化終了文字列として、“<SECTION>”と“<SUBSECTION>”が設定されているものとする。

【0104】この場合、最初に現れる符号化開始文字列は、“</SECTION>”（2行目）であるので、文書データの先頭から2行目の“</SECTION>”までの各文字は、非圧縮データ出力処理ループで処理される。そして、“</SECTION>”の次の文字から圧縮データ出力処理ループによる処理が開始されることになる。圧縮データ出力処理ループの開始後、最

初に現れる符号化終了文字列は、“<SECTION>”（4行目）である。このため、3行目の始めから、4行目の“<SECTION>”までの各文字は符号化されて出力される。そして、“<SECTION>”内の“>”の符号化が終わった際に、符号表の初期化が行われるとともに、“<SECTION>”が圧縮文書ファイルとインデックスファイルに書き込まれる。

【0105】このような一連の動作が、圧縮文書ファイル内の各データに対して繰り返される結果、第3実施形態の文書管理装置では、図13、図14にそれぞれ示したような圧縮文書ファイルとインデックスファイルが作成されることになる。

【0106】すなわち、第3実施形態の文書管理装置が作成する圧縮文書ファイルには、第1実施形態の文書管理装置が作成する圧縮文書ファイル（図5）内の各非圧縮データに、符号化終了文字列（開始タグ）を付加した非圧縮データが記憶される。そして、インデックスファイルには、圧縮文書ファイル内の非圧縮データと同じデータが記憶される。また、圧縮データ出力処理ループの終了時に、符号表の初期化が行われているので、圧縮文書ファイル内の各圧縮データは、単独で復元できるものとなっている。

【0107】以下、圧縮文書ファイルの内容の指定した範囲だけを復元させる処理であるインデックス対応領域復元処理の詳細を説明する。図15に、インデックス対応領域復元処理時の文書管理装置（データ処理装置）の動作手順を示す。なお、この図に示した流れは、ユーザから、文書データの特定情報を含む所定の指示が与えられた場合に開始される。

【0108】図示したように、文書管理装置（データ処理装置）は、ユーザから所定の指示を受けた場合、その指示で指定された文書データに応じたインデックスファイルの内容を表示装置に表示する（ステップS601）。なお、このステップにおいて、データ処理装置は、インデックスファイル内の、開始および終了タグで挟まれたデータ（以下、インデックスと表記する）だけを表示装置に表示している。例えば、図14に示したインデックスファイルに対応する文書データが処理対象として指示されていた場合、表示装置には、図16に示したようなデータが表示される。

【0109】その後、データ処理装置は、ユーザの指示入力を待機する状態に移行する（ステップS602）。ステップS602において、データ処理装置は、画面上で出力対象のインデックスを指定するための処理であるマウスのクリックが行われるのを待機しており、ユーザは、マウスを操作することによってデータ処理装置に対して実行すべき処理を指示する。なお、このステップにおいて、ユーザは、他のインデックスファイルの内容表示を行わせるための指示や、インデックスファイルの内容表示を終了させるための指示が入力できるのである

が、ここでは、いずれかのインデックス上にマウスカーソルが位置している状態で、マウスがクリックされた場合の動作だけを説明することにする。

【0110】いずれかのインデックス上にマウスカーソルが位置している状態で、マウスがクリックされた場合（ステップS602；Y）、データ処理装置は、そのインデックスが選択されたことを認識し、インデックスファイルを参照することによって、選択されたインデックスに対応するインデックスデータ（タグで挟まれたインデックス）を特定する（ステップS603）。

【0111】そして、データ処理装置は、特定したインデックスデータが、“TITLE”に関するものであるか否かを判断し、“TITLE”に関するものであった場合（ステップS604；Y）には、対象となっている文書データに対応する圧縮文書ファイルの内容を全て復元する処理である全体復元処理を実行（ステップS605）し、復元結果を表示あるいはファイルとして記憶して、処理を終了する。

【0112】図17に、全体復元処理時のデータ処理装置の動作手順を示す。なお、この処理は、圧縮文書ファイルを復元することが指示された際にも実行される。図示したように、全体復元処理時、データ処理装置内では、非圧縮データ処理ループ（ステップS701～S703）が実行される。非圧縮データ処理ループ実行時、データ処理装置は、まず、圧縮文書ファイル内の最初の1文字分のデータがそのまま復元結果として出力する（ステップS701）。次いで、圧縮文書ファイル内の全てのデータに対する処理が終了しているか否かを判断する（ステップS702）。そして、処理すべきデータが残っていた場合（ステップS702；N）には、処理した文字列（そのときに処理した文字を含む）が、いずれかの復号開始文字列と一致しているか否かを判断する（ステップS703）。

【0113】処理した文字列が各復号開始文字列と一致していなかった場合（ステップS703；N）、データ処理装置は、ステップS701からの処理を再度実行する。一方、処理した文字列が復号開始文字列の1つと一致していた場合（ステップS703；Y）、データ処理装置は、圧縮データ処理ループ（ステップS704～S707）を開始する。

【0114】圧縮データ処理ループにおいて、データ処理装置は、まず、圧縮文書ファイルのデータ（符号）を必要量読み込み、その符号の復号結果である文字を出力する（ステップS704）。なお、このステップにおける復号は、既に復号を終えた文字列（文脈）を参照した形で行われる。そして、データ処理装置は、復号に用いた文脈に関する符号表の内容を更新する（ステップS705）。

【0115】次に、データ処理装置は、圧縮文書ファイル内の全てのデータに対する処理が終了しているか否か

を判断する(ステップS706)。そして、処理すべきデータが残っていた場合(ステップS706;N)には、復号を終えた文字列が、いずれかの復号終了文字列と一致するか否かを判断する(ステップS707)。そして、いずれの復号終了文字列とも一致していなかった場合(ステップS707;N)、データ処理装置は、ステップS704からの処理を開始する。一方、復号を終えた文字列が、復号終了文字列の1つと一致していた場合(ステップS707;Y)、データ処理装置は、全ての文脈に関する符号表を初期化(ステップS708)する。次いで、データ処理装置は、圧縮文書ファイル内の、次に処理すべきデータの先頭部分に存在している、復号終了文字列を読み飛ばす(ステップS709)。すなわち、圧縮文書ファイル作成時に付加した符号化終了文字列を読み飛ばす。その後、データ処理装置は、非圧縮データ処理ループ(ステップS701~S703)を開始する。

【0116】データ処理装置は、このような処理を、圧縮文書ファイル内の全てのデータに対する行った(ステップS706;Y)に、全体復元処理を終了する。図15に戻って、インデックス対応領域復元処理の説明を続ける。

【0117】インデックスデータが、“TITLE”に関するものでなかった場合(ステップS604;N)、データ処理装置は、そのインデックスデータの先頭のタグを終了制御文字列として取得(記憶)する(ステップS606)。そして、圧縮文書ファイルの内容のうち、選択されたインデックスに関係するデータだけを復元する処理である部分復元処理を実行(ステップS607)し、処理を終了する。

【0118】図18に、部分復元処理時のデータ処理装置の動作の流れを示す。部分復元処理の全体的な流れは、全体復元処理(図17)と同じであり、開始条件と終了条件だけが異なっている。このため、ここでは、異なる部分に関する説明だけを行うことにする。

【0119】全体復元処理では、圧縮文書ファイルの先頭から復元処理が開始される。これに対して、部分復元処理では、最初に、インデックスデータを基に復元開始位置が特定される(ステップS800)。すなわち、圧縮文書ファイルの中から、選択されたインデックスに応じたインデックスデータが検索され、検索されたインデックスデータの最初の文字が復元開始位置として特定される。

【0120】そして、その復元開始位置からのデータが、全体復元処理と同様の手順で処理されていく。また、全体復元処理では、圧縮文書ファイル内の全てのデータに関する処理が完了したときに、処理が終了される。これに対して、部分復元処理では、符号表の初期化(ステップS808)後に、終了判定(ステップS809)が行われる。具体的には、データ処理装置は、ス

ップS807で見い出した復号終了文字列が、装置内に記憶されている終了制御文字列と一致しているか否かを判断する。そして、一致していなかった場合(ステップS809;N)には、全体復元処理と同様に、次に処理すべき部分に存在している復号終了文字列を読み飛ばして(ステップS810)、非圧縮データ処理ループを開始する。

【0121】一方、復号終了文字列と終了制御文字列が一致していた場合(ステップS809;Y)には、復元結果から、終了制御文字列を取り除いて(ステップS811)、部分復元処理を終了する。

【0122】以下、図14の“2. 特許請求の範囲”が指定された場合を例に、インデックス対応領域復元処理をさらに具体的に説明する。この場合、対応するインデックスデータは、“<SECTION>2. 特許請求の範囲</SECTION>”であるので、終了制御文字列として“<SECTION>”が特定される。そして、部分復元処理が開始され、まず、圧縮文書ファイル内から“<SECTION>2. 特許請求の範囲</SECTION>”が検索される。次いで、圧縮文書ファイル内の検索された文字列の最初の文字から復元が開始され、“<SECTION>2. 特許請求の範囲</SECTION>”が非圧縮データ処理ループによって処理されることになる。その後に行われる最初の圧縮データ処理ループでは、圧縮文書ファイル内に記憶された圧縮データである復号終了文字列“<SUBSECTION>”が復元される。しかし、その文字列は、終了制御文字列“<SECTION>”と一致していないので、データ処理装置は、圧縮文書ファイルの復元を続ける。

そして、次に圧縮データ処理ループを実行した際には、“<SECTION>”が復元されるので、データ処理装置は、その“<SECTION>”を復元結果から取り除き、部分復元処理を終える。すなわち、“<SECTION>3. 発明の詳細な説明</SECTION>”の前の部分まで復元を行い、部分復元処理を終了する。

【0123】結局、インデックス対応領域復元処理では、図19に模式的に示したように、選択されたインデックスに応じた領域(図中、野線で囲んだ領域)内のデータが復元される。すなわち、タイトルに関するインデックスを選択した場合には、全ての内容が復元され、サブセクションレベルのインデックスを選択した場合には、そのサブセクションレベルのデータだけが復元される。また、セクションレベルのインデックスを選択した場合、そのセクションに関するデータ(サブセクションレベルのデータを含む)が全て復元される。

【0124】このように、第3実施形態の文書管理装置によれば、圧縮文書ファイルの一部分だけを復元することができる。以上説明したように、第3実施形態の文書管理装置では、圧縮文書ファイル内の各非圧縮データ

に、符号化終了文字列（開始タグ）を含ませるために、圧縮データの出力後に符号化終了文字列を付加するといった手順を採用している。しかし、処理すべき文字を何文字がバッファリングしておき、開始タグの一部ではないことが確定した文字に対して符号化が行われるように装置を構成することによって、圧縮文書ファイル内の各非圧縮データに、開始タグを含ませることも可能である。ただし、このように装置を構成した場合、圧縮文書ファイル内の圧縮データに対しては、開始タグを検索しつつ（予め圧縮データ、非圧縮データの境を定めておき）、復号を行うことになる。

【0125】第4実施形態

第4実施形態の文書管理装置は、第3実施形態の文書管理装置と同じインデックスファイルを作成する。ただし、第4実施形態の文書管理装置は、静的符号化を用いて圧縮された第1圧縮データと、動的符号化により圧縮された第2圧縮データが混在する圧縮文書ファイルを作成する。また、圧縮文書ファイルとインデックスファイルを関係づけるファイルとして、対応領域管理ファイルを作成する。

【0126】図20に、第4実施形態の文書管理装置（データ処理装置）による圧縮文書ファイル作成手順を示す。なお、第4実施形態の文書管理装置では、符号化開始文字列として“</TITLE>”、“</SECTION>”、“</SUBSECTION>”が与えられており、符号化終了文字列として、“<SECTION>”、“<SUBSECTION>”が与えられている。

【0127】文書データの圧縮を最初に指示された際、データ処理装置内では、第1圧縮データ出力処理ループ（ステップS901～S903）が開始される。第1圧縮データ出力処理ループ実行時、データ処理装置は、まず、文書データ内の1文字（対象文字）をそのままインデックスファイルに出力するとともに、その対象文字を、静的符号表を用いて符号化することによって得られた符号を、圧縮文書ファイル内に書き込む（ステップS901）。なお、データ処理装置は、このステップにおいて、圧縮文書ファイルに対して出力したデータサイズの積算も行う。

【0128】次いで、データ処理装置は、文書データを構成する全ての文字に対する処理が終了しているか否かを判断する（ステップS902）。そして、処理すべきデータ（文字）が残っていた場合（ステップS902；N）には、処理した文字列が、予め定められている符号化開始文字列の1つと一致している否かを判断する（ステップS903）。

【0129】処理した文字列が符号化開始文字列と一致していなかった場合（ステップS903；N）、データ処理装置は、ステップS901からの処理を再度実行する。一方、処理した文字列が、符号化開始文字列の1つ

と一致した場合（ステップS903；Y）、データ処理装置は、第2圧縮データ出力処理ループ（ステップS904～S907）を開始する。

【0130】第2圧縮データ出力処理ループ実行時、データ処理装置は、文書データから次の1文字を読み込み、圧縮文書ファイル内に、その対象文字に対応する符号を出力する（ステップS904）。なお、このステップにおける符号出力は、対象文字の文脈を参照した形で行われる。また、データ処理装置は、このステップにおいて、圧縮文書ファイルに書き込んだデータサイズの積算も行う。次いで、データ処理装置は、符号化に用いた文脈に関する符号表の内容を更新する（ステップS905）。

【0131】次に、データ処理装置は、文書データを構成する全ての文字に対する処理が終了しているか否かを判断する（ステップS906）。そして、処理すべきデータが残っていた場合（ステップS906；N）、データ処理装置は、処理した文字列が、予め定められている符号化終了文字列の1つと一致するか否かを判断する

（ステップS907）。そして、処理した文字列がいずれの符号化終了文字列と一致していなかった場合（ステップS907；N）、データ処理装置は、ステップS904からの処理を再度実行する。一方、処理した文字列が、符号化終了文字列の1つと一致していた場合（ステップS907；Y）、データ処理装置は、符号表の初期化を行う（ステップS908）。

【0132】次いで、データ処理装置は、ステップS907で検出した符号化終了文字列をインデックスファイルに出力するとともに、その文字列を静的符号化した符号を圧縮文書ファイルに出力する（ステップS909）。また、データ処理装置は、格納した静的符号の、圧縮文書ファイル内での格納位置情報（静的符号の先頭ビットまでの圧縮文書ファイル内のデータサイズ）を、対応関係管理ファイルに記憶する（ステップS910）。なお、データ処理装置は、それまで積算してきたデータサイズの積算結果を基に格納位置情報を定め、格納位置情報を定めた後に、積算結果に、ステップS909で書き込んだ静的符号のデータサイズを積算する。

【0133】その後、データ処理装置は、第1圧縮データ出力処理ループを再度実行する。そして、データ処理装置は、第1圧縮データ出力処理ループにおいて全データに関する処理が終了したことを検出した際（ステップS902；Y）、あるいは、第2圧縮データ出力処理ループにおいて全データに関する処理が終了したことを検出した際（ステップS906；Y）に、圧縮文書ファイル作成処理を終了する。

【0134】すなわち、第4実施形態の文書管理装置では、図21に模式的に示したような、静的符号化による第1圧縮データ（図中、下線を付した部分）と、動的符号化による第2圧縮データが混在する圧縮文書ファイル

が作成される。そして、2番目以降の第1圧縮データの先頭ビットの格納位置が記憶された対応関係管理ファイルが作成される。

【0135】次に、第4実施形態の文書管理装置におけるインデックス対応領域復元処理を説明する。インデックス対応領域復元処理の全体的な流れは、図15に示したものと同一であるので、説明は省略する。

【0136】図22に、第4実施形態の文書管理装置における部分復元処理の流れを示す。この部分復元処理の基本的な流れは、既に説明した第3実施形態の文書管理装置による部分復元処理と同じものとなっている。このため、ここでは、動作内容が異なるステップだけを説明することにする。

【0137】第3実施形態の文書管理装置では、インデックスデータの格納位置を検索することによって、復元開始位置が特定される。これに対して、第4実施形態の文書管理装置では、対応関係管理ファイルを参照することによって、復元開始位置が特定（ステップS1000）される。具体的には、データ処理装置は、まず、ユーザによって指定されたインデックスデータがインデックスファイル内の何番目のデータであるかを判別する。例えば、M番目のデータであった場合、データ処理装置は、対応関係管理ファイル内の、M-1番目の格納位置情報を読み出す。そして、その格納位置情報によって定められる位置を、復元開始位置と特定する。

【0138】その後、復元開始位置以降のデータに対して処理が行われていくことになるが、第4実施形態の文書管理装置では、インデックスに関する処理時に、静的符号表を用いた復号が行われる。

【0139】すなわち、復元開始位置の特定の直後に行われるループでは、その最初に、圧縮文書ファイルから必要量のデータを読み出し、そのデータを静的符号表を用いて復号する処理が行われる（ステップS1001）。また、ステップS1110では、復号終了文字列に対応する符号が読み飛ばされる。

【0140】

【発明の効果】以上、詳細に説明したように、本発明の文書管理装置によれば、キーワード検索が可能な形態で文書データが圧縮されてファイル化される。このため、本発明の文書管理装置によれば、ハードディスク装置などのファイルを記憶するための装置の記憶容量を有効に活用しつつ、高速な文書データ処理が行えることになる。

【0141】また、本発明のデータ圧縮方法によれば、キーワード検索が可能な形態でデータを圧縮することができる。そして、本発明のデータ復元方法によれば、本発明のデータ圧縮方法によって圧縮されたデータを復元できる。

【図面の簡単な説明】

【図1】本発明の第1実施形態の文書管理装置の構成を

示すブロック図である。

【図2】第1実施形態の文書管理装置による圧縮文書ファイル作成手順を説明するための機能ブロック図である。

【図3】第1実施形態の文書管理装置による圧縮文書ファイル作成手順を示す流れ図である。

【図4】SGML形式で記述された文書データの一例を示した図である。

【図5】図4に示した文書データから、第1実施形態の文書管理装置によって作成される圧縮文書ファイルの概要を示す図である。

【図6】本発明の第1実施形態の文書管理装置の復元動作を説明するための機能ブロック図である。

【図7】本発明の第1実施形態の文書管理装置による圧縮文書ファイル復元手順を示す流れ図である。

【図8】本発明の第2実施形態の文書管理装置による圧縮文書ファイルの作成手順を説明するための機能ブロック図である。

【図9】本発明の第2実施形態の文書管理装置による圧縮文書ファイルの作成手順を示す流れ図である。

【図10】本発明の第2実施形態の文書管理装置による圧縮文書ファイルの復元手順を説明するための機能ブロック図である。

【図11】本発明の第2実施形態の文書管理装置による圧縮文書ファイルの復元手順を示す流れ図である。

【図12】本発明の第3実施形態の文書管理装置による圧縮文書ファイルの作成手順を示す流れ図である。

【図13】第3実施形態の文書管理装置によって作成される圧縮文書ファイルの概要図である。

【図14】第3実施形態の文書管理装置によって作成されるインデックスファイルの概要図である。

【図15】第3実施形態の文書管理装置におけるインデックス対応領域復元処理の流れ図である。

【図16】第3実施形態の文書管理装置による圧縮文書ファイルの作成手順を示す流れ図である。

【図17】第3実施形態の文書管理装置において実行される全体復元処理の流れ図である。

【図18】第3実施形態の文書管理装置において実行される部分復元処理の流れ図である。

【図19】インデックス対応領域復元処理において復元される領域と、インデックスとの対応関係を示した説明図である。

【図20】本発明の第4実施形態の文書管理装置による圧縮文書ファイル作成手順を示す流れ図である。

【図21】第4実施形態の文書管理装置によって作成される圧縮文書ファイルの概要を示す図である。

【図22】第4実施形態の文書管理装置において実行される部分復元処理の流れ図である。

【符号の説明】

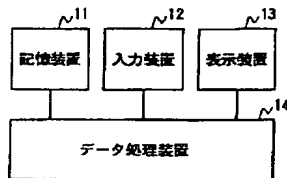
11 記憶装置

12 入力装置
 13 表示装置
 14 データ処理装置
 101、201 第1文字列保持部
 102、202 第2文字列保持部
 103、203 入力文字列保持部
 104、204 文脈保持部
 105 符号化開始文字列検索部
 106 符号化終了文字列検索部
 107、207 スイッチ
 108、208 符号保持部

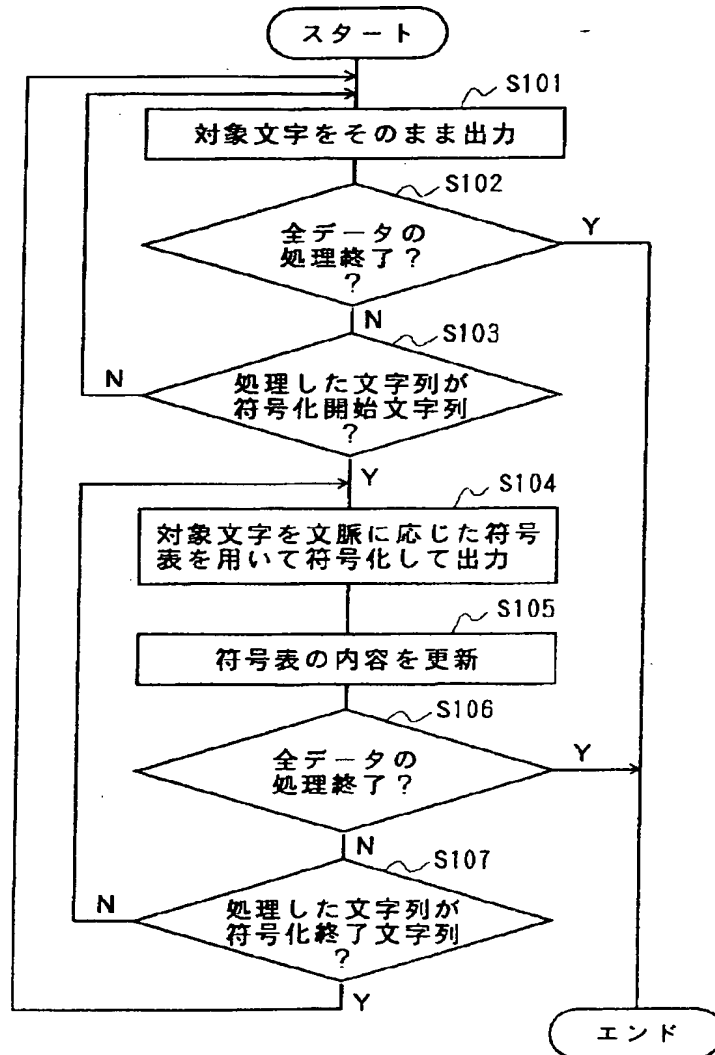
* 109 符号化部
 110、210 符号更新部
 122 置換部
 123 置換表保持部
 130、230 入力端子
 131、231 出力端子
 205 復号開始文字列検索部
 206 復号終了文字列検索部
 209 復号部
 10 222 逆置換部
 * 223 逆置換表保持部

【図1】

第1実施形態の文書管理装置のブロック図

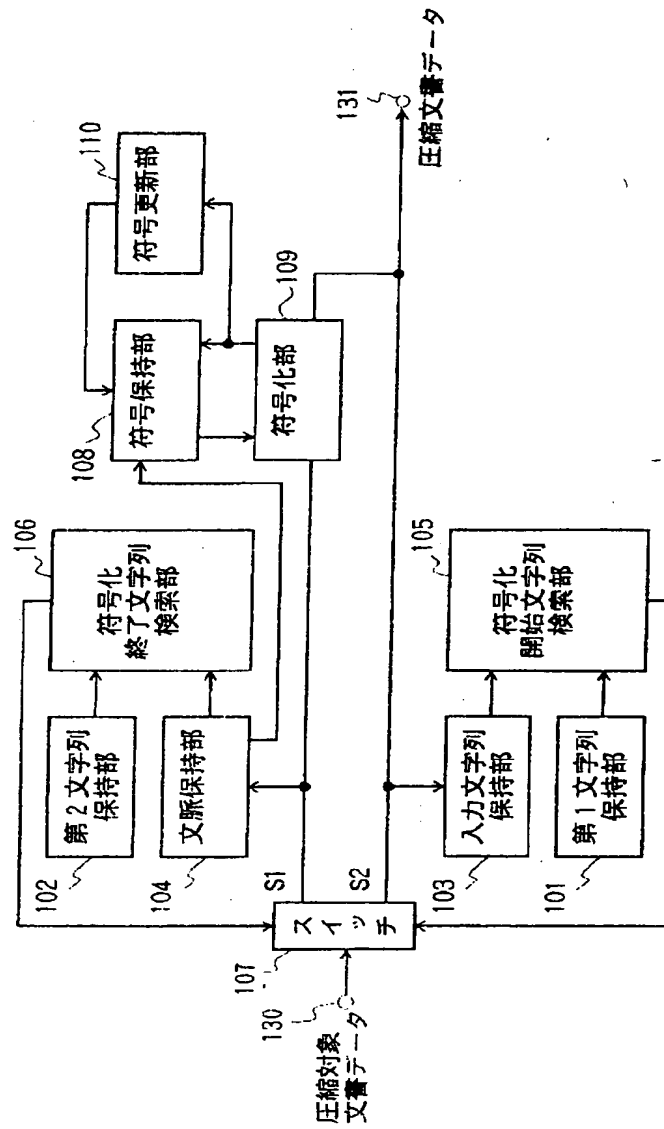


【図3】

第1実施形態の文書管理装置の
圧縮文書ファイル作成手順を示す流れ図

【図2】

第1実施形態の文書管理装置における
圧縮文書ファイル作成手順を説明するための機能ブロック図



【図4】

```

<TITLE>発明説明書</TITLE><P>
<SECTION>1. 発明の名称</SECTION>
  文書管理装置および文書管理方法
<SECTION>2. 特許請求の範囲</SECTION>
<SUBSECTION>タグ付き文書データを扱う装置であって、</SUBSECTION>
<LIST>
  <ITEM> .....
</LIST>を備えることを特徴とする文書管理装置。<P>
<SECTION>3. 発明の詳細な説明</SECTION>
<SUBSECTION>(1)産業上の利用分野</SUBSECTION>
<PARAGRAPH>本発明は、文書管理装置および文書管理方法に関し、...
<SUBSECTION>(2)従来の技術</SUBSECTION>
<PARAGRAPH>近年、文字コード、ベクトル情報、画像など様々な種類のデータ
がコンピュータで扱われるようになっている。...
<PARAGRAPH>一方、最近では計算機で扱う文書の形式を統一する動きがある。
これまで、計算機あるいはアプリケーションによってばらばらであった文書の
形式を異なる計算機環境でも使用できるようにするものである。<TT>SG
ML</TT>は、...

```

【図5】

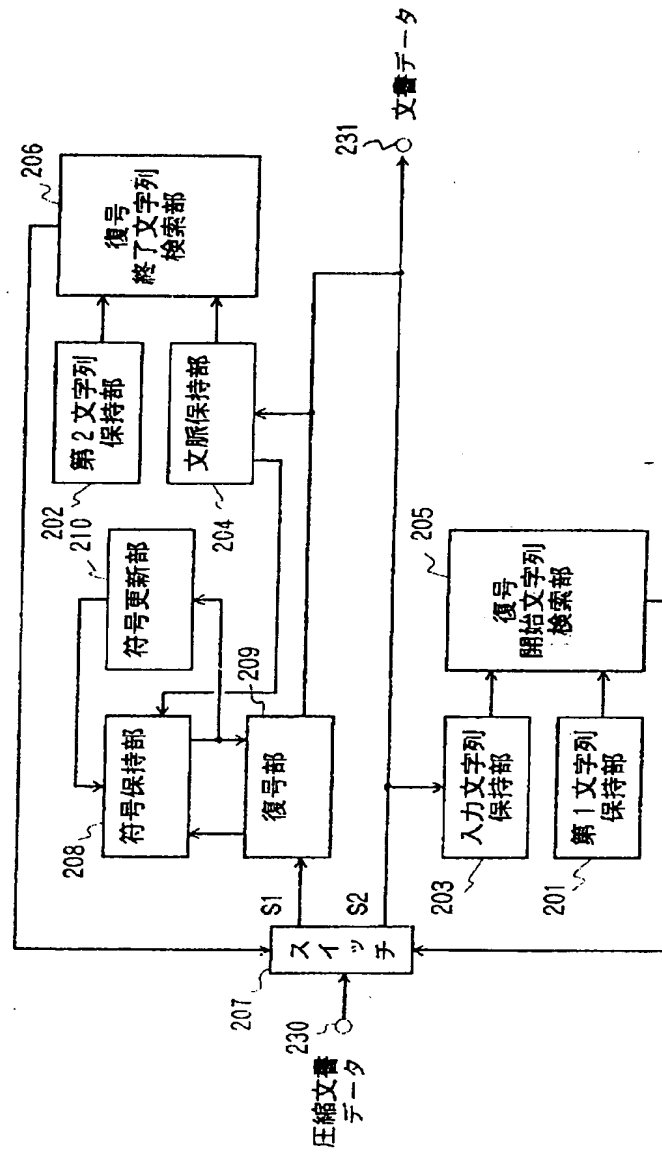
```

<TITLE>発明説明書</TITLE><P>
<SECTION>1. 発明の名称</SECTION>
  275fe5a..... (圧縮データ)
2. 特許請求の範囲</SECTION>
  61fdc... (圧縮データ)
  タグ付き文書データを扱う装置であって、</SUBSECTION>
  6ef208c..... (圧縮データ)
3. 発明の詳細な説明</SECTION>
  23fdc... (圧縮データ)
(1)産業上の利用分野</SUBSECTION>
  425fea..... (圧縮データ)
(2)従来の技術</SUBSECTION>
  e75f5a..... (圧縮データ)

```

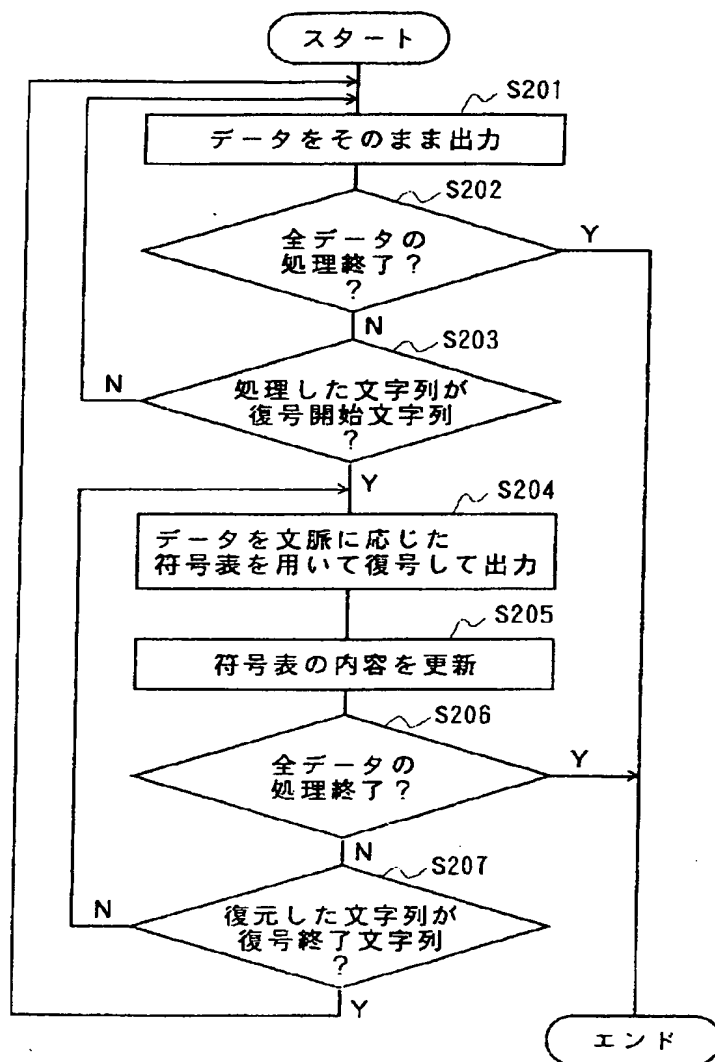
【図6】

第1実施形態の文書管理装置における
圧縮文書ファイル復元手順を説明するための機能ブロック図



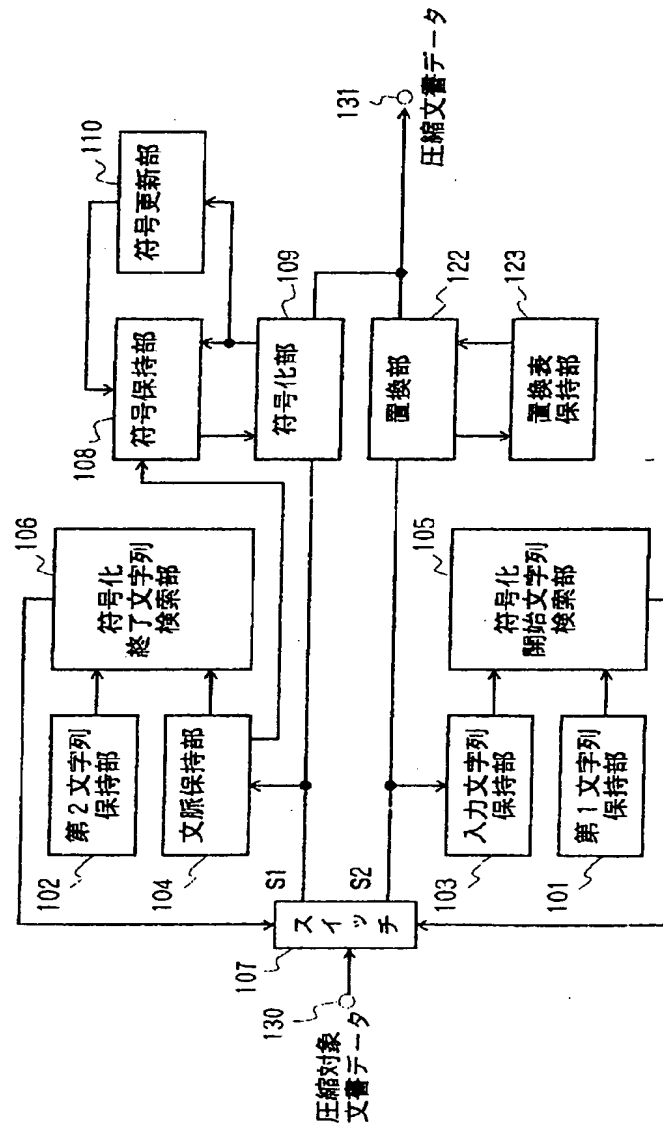
【図7】

第1実施形態の文書管理装置の
圧縮文書ファイル復元手順を示す流れ図



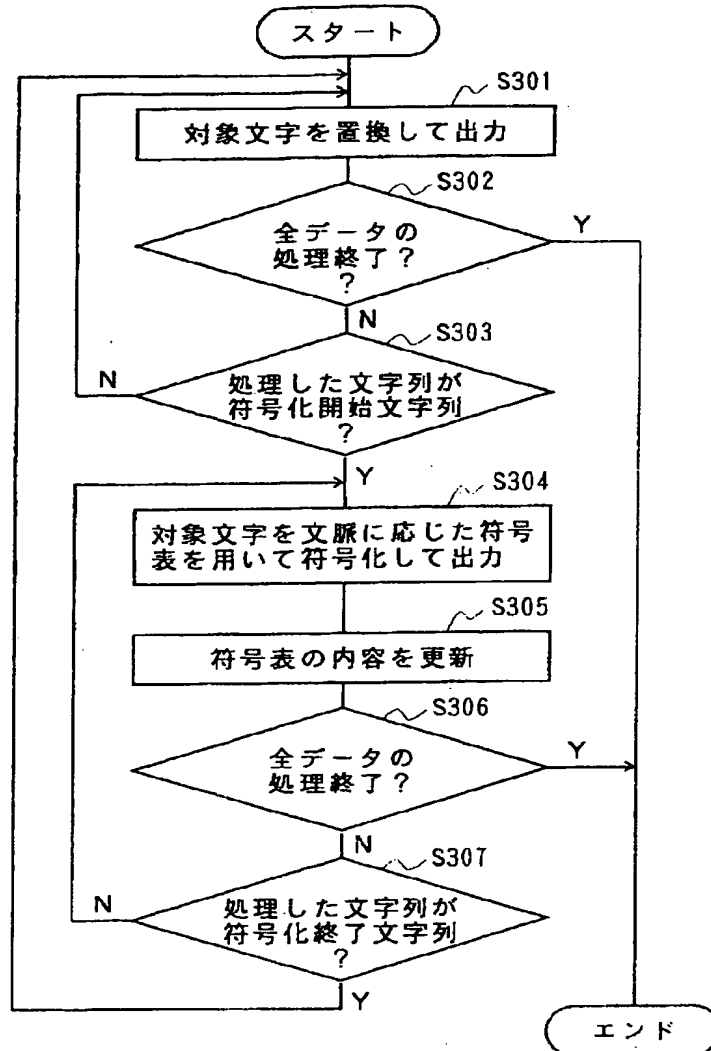
【図8】

第2実施形態の文書管理装置における
圧縮文書ファイル作成手順を説明するための機能ブロック図



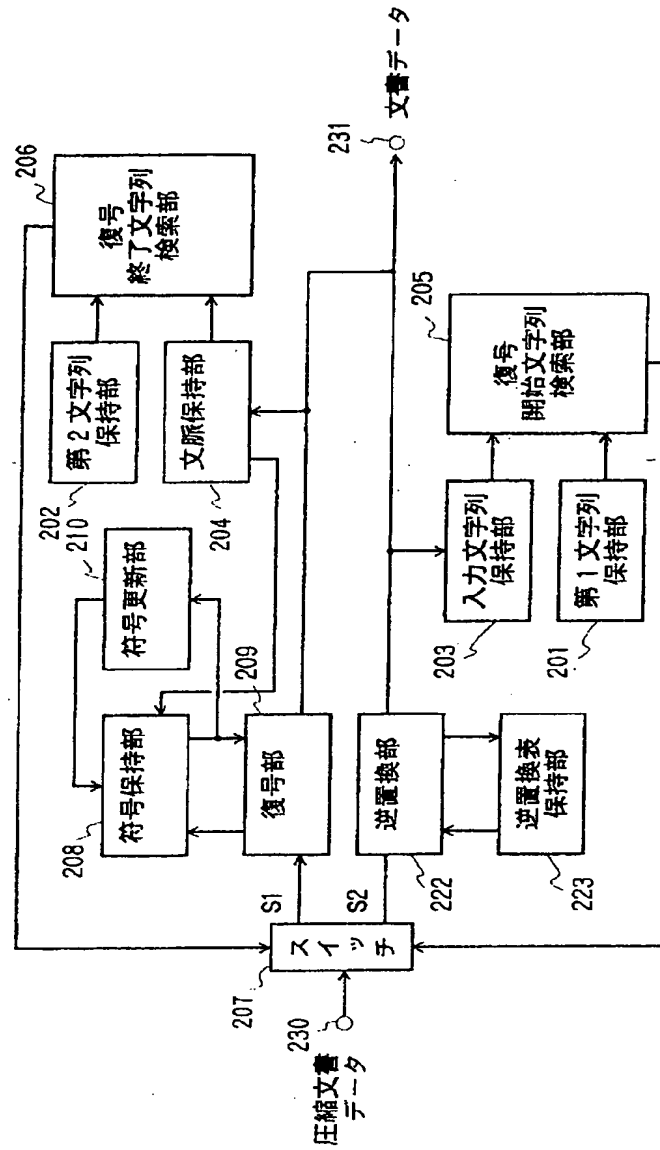
【図9】

第2実施形態の文書管理装置の
圧縮文書ファイル作成手順を示す流れ図



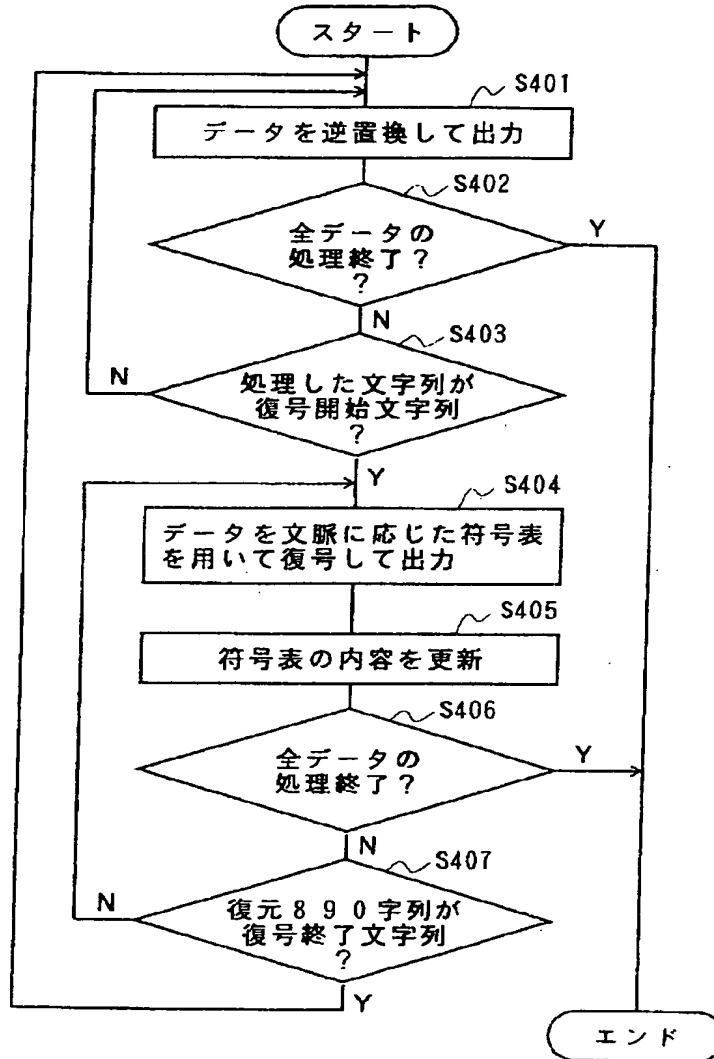
【図10】

第2実施形態の文書管理装置における
圧縮文書ファイル復元手順を説明するための機能ブロック図

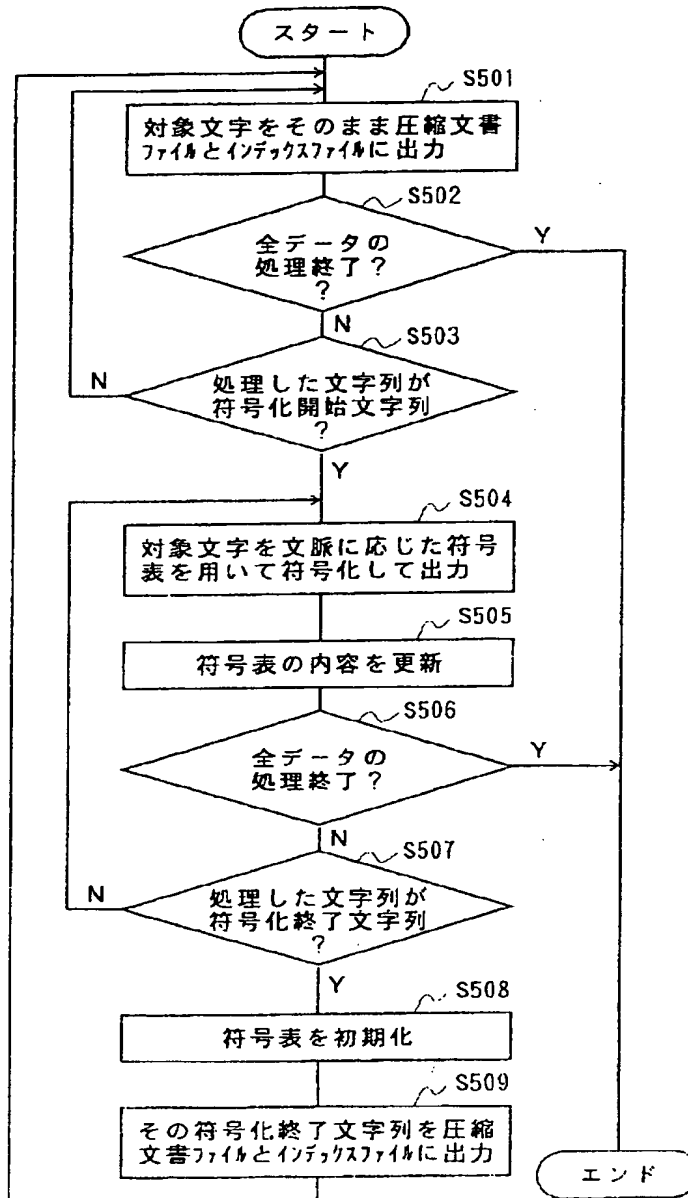


【図11】

第2実施形態の文書管理装置の
圧縮文書ファイル復元手順を示す流れ図



【図12】

第3実施形態の文書管理装置の
圧縮文書ファイル作成手順を示す流れ図

【図13】

第3実施形態の文書管理装置が作成する圧縮文書ファイルの概要図

```

<TITLE>発明説明書</TITLE><P>
<SECTION>1. 発明の名称</SECTION>
  275fe5a..... (圧縮データ)
<SECTION>2. 特許請求の範囲</SECTION>
  61fdc...      (圧縮データ)
<SUBSECTION>タグ付き文書データを扱う装置であって、</SUBSECTION>
  6ef208c..... (圧縮データ)
<SECTION>3. 発明の詳細な説明</SECTION>
  23fdc...      (圧縮データ)
<SUBSECTION>(1)産業上の利用分野</SUBSECTION>
  425fea..... (圧縮データ)
<SUBSECTION>(2)従来技術</SUBSECTION>
  e75f5a..... (圧縮データ)

```

【図14】

第3実施形態の文書管理装置が作成するインデックスファイルの概要図

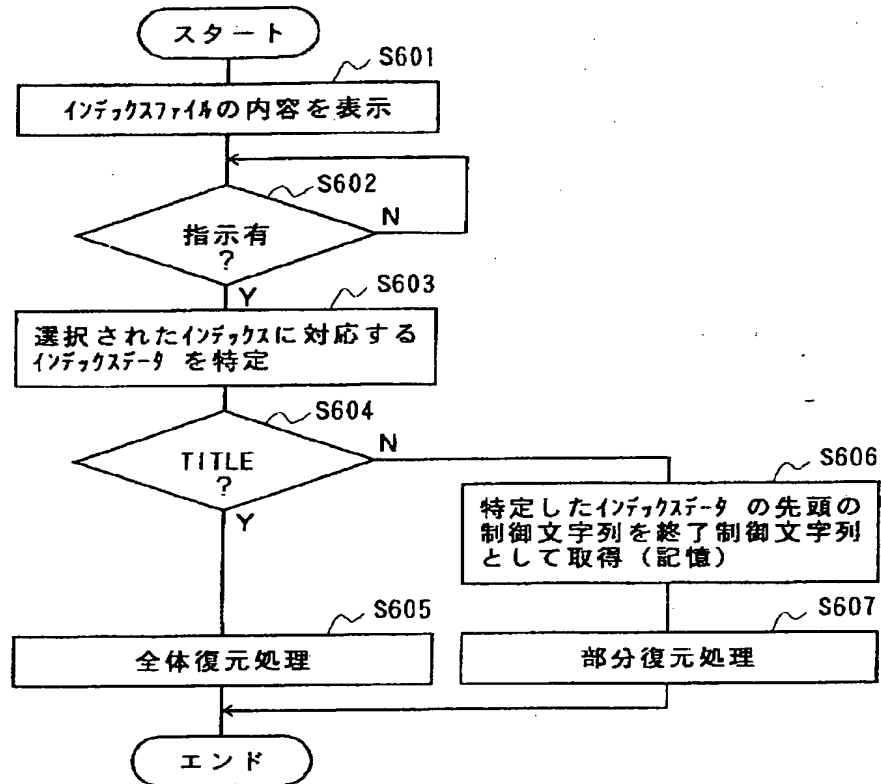
```

<TITLE>発明説明書</TITLE><P>
<SECTION>1. 発明の名称</SECTION>
<SECTION>2. 特許請求の範囲</SECTION>
<SUBSECTION>タグ付き文書データを扱う装置であって、</SUBSECTION>
<SECTION>3. 発明の詳細な説明</SECTION>
<SUBSECTION>(1)産業上の利用分野</SUBSECTION>
<SUBSECTION>(2)従来技術</SUBSECTION>

```

【図15】

インデックス対応領域復元処理の流れ図



【図16】

インデックス対応領域復元処理において
表示装置に表示される内容を示した説明図

発明説明書

1. 発明の名称

文書管理装置および文書管理方法

2. 特許請求の範囲

タグ付き文書データを扱う装置であって、

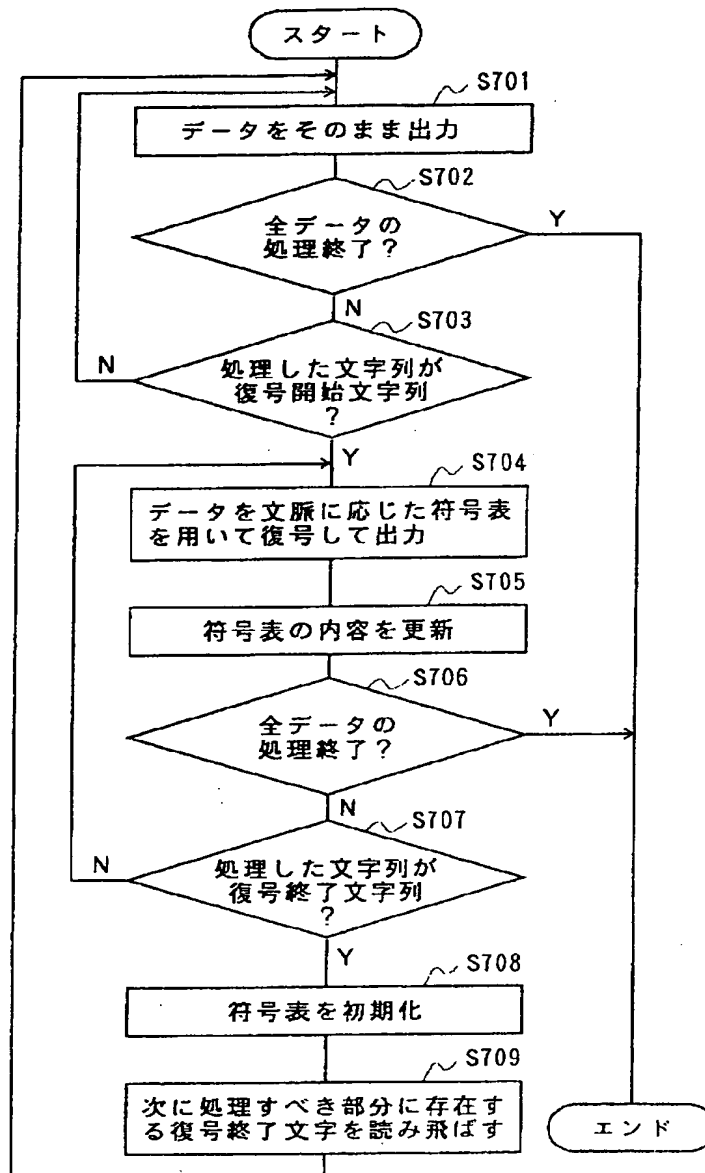
3. 発明の詳細な説明

(1)産業上の利用分野

(2)従来技術

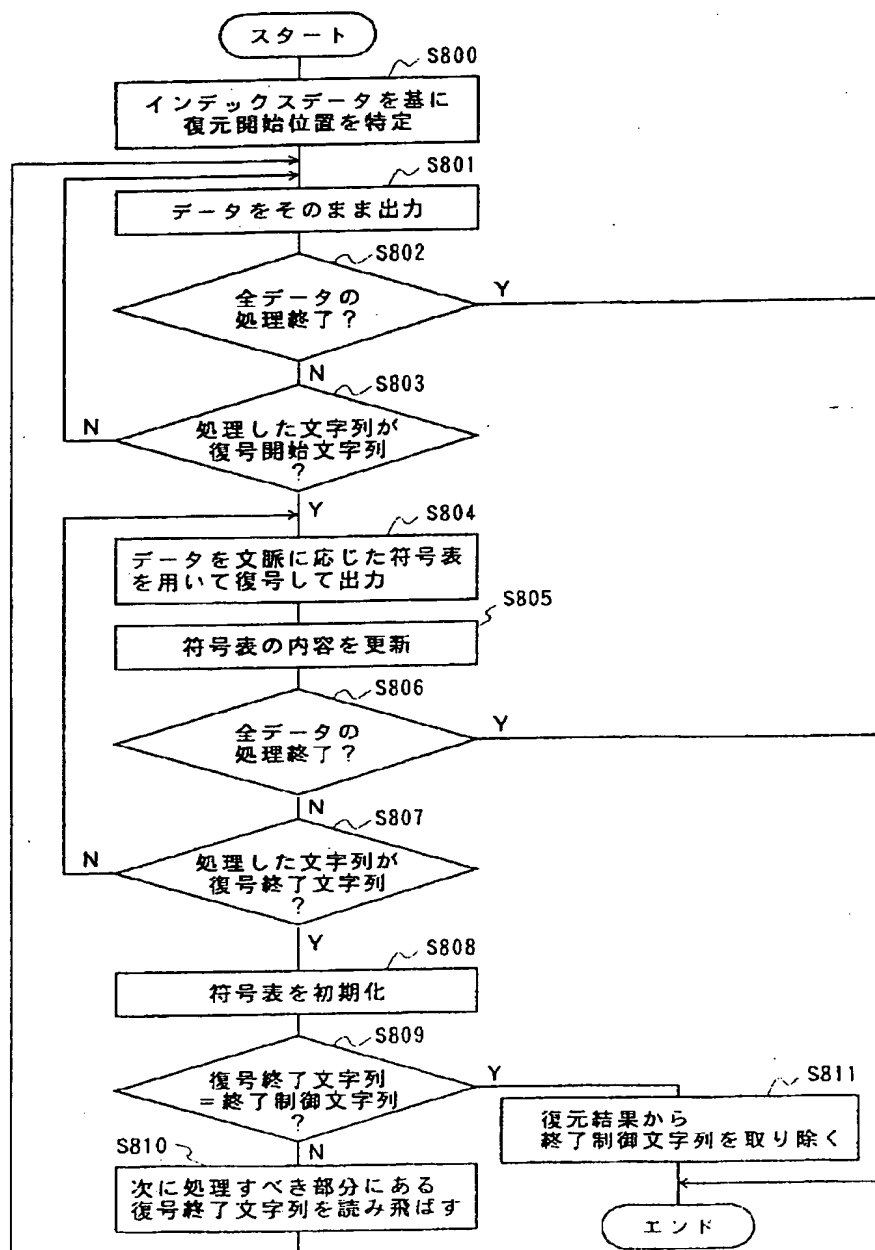
【図17】

第3実施形態の文書管理装置において
実行される全体復元処理の流れ図



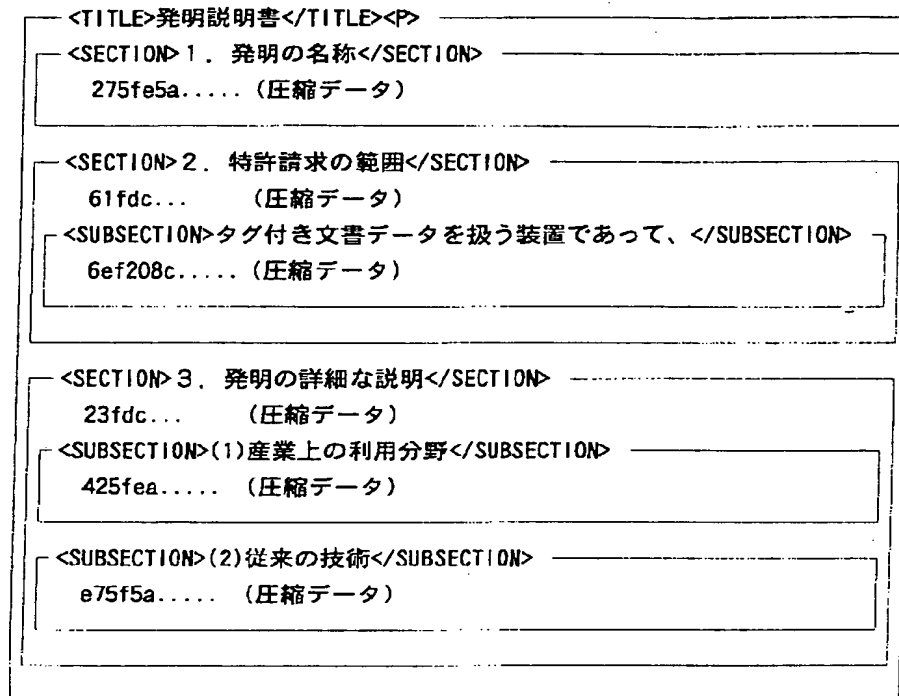
【図18】

第3実施形態の文書管理装置において実行される部分復元処理の流れ図



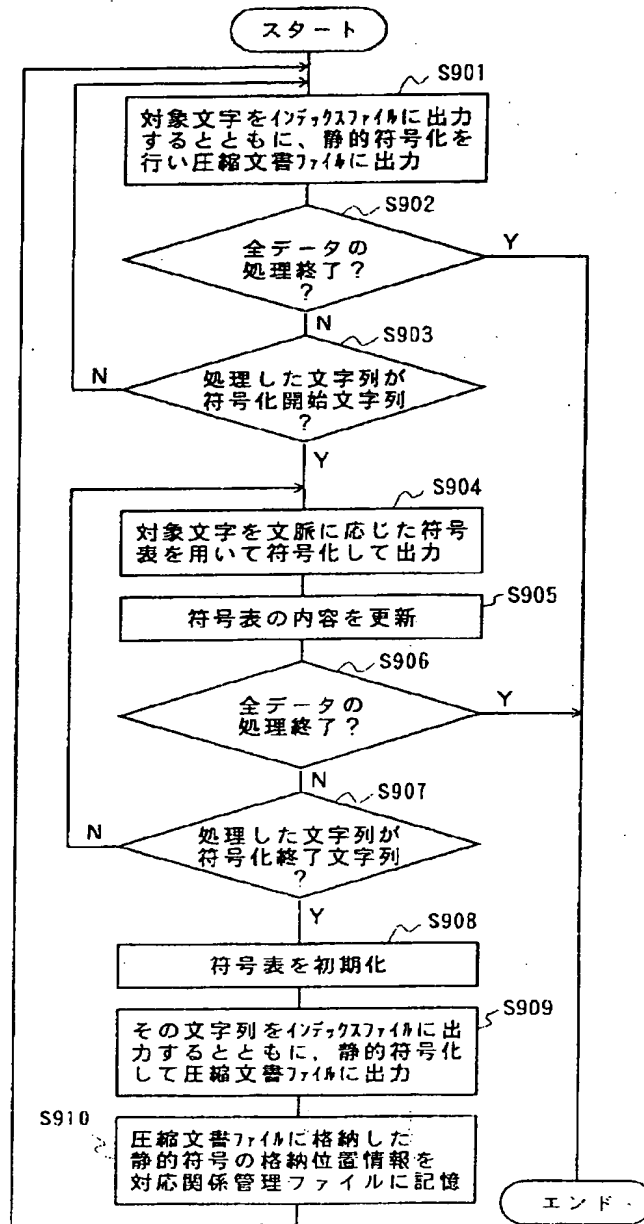
【図19】

インデックス対応領域復元処理において、復元される領域と
インデックスとの対応関係を示した説明図。



【図20】

第4実施形態の文書管理装置の
圧縮文書ファイル作成手順を示す流れ図



【図21】

第4実施形態の文書管理装置が作成する圧縮文書ファイルの概要図

<TITLE>発明説明書</TITLE>
2f... (第2圧縮データ)
<SECTION>1. 発明の名称</SECTION>
275fe5a..... (第2圧縮データ)
<SECTION>2. 特許請求の範囲</SECTION>
61fdc... (圧縮データ)
<SUBSECTION>タグ付き文書データを扱う装置であって、</SUBSECTION>
6ef208c..... (第2圧縮データ)
<SECTION>3. 発明の詳細な説明</SECTION>
23fdc... (第2圧縮データ)
<SUBSECTION>(1)産業上の利用分野</SUBSECTION>
425fea..... (第2圧縮データ)
<SUBSECTION>(2)従来の技術</SUBSECTION>
e75f5a..... (第2圧縮データ)

【図22】

第4実施形態の文書管理装置において実行される部分復元処理の流れ図

